2008

REDUCING BACTERIAL RESISTANCE THROUGH BETTER ANTIBIOTIC PRESCRIPTION PRACTICES

Christine Ouma
Virginia Commonwealth University

College of Humanities & Sciences
Virginia Commonwealth University

This is to certify that the thesis prepared by Christine Ouma entitled

REDUCING BACTERIAL RESISTANCE THROUGH BETTER ANTIBIOTIC PRESCRIPTION PRACTICES

has been approved by her committee as satisfactory completion of the thesis requirement for the degree of Master of Science in Statistics

Dr. James E. Mays, Department of Statistical Sciences & Operations Research

Dr. Edward Boone, Department of Statistical Sciences & Operations Research

Dr. Amy Pakyz, School of Pharmacy

Dr. D'Arcy P. Mays, Chair of the Department of Statistical Sciences & Operations Research

Dr. Robert Holsworth, Dean of the College of Humanities & Sciences

Dr. F. Douglas Boudinot, Dean of the School of Graduate Studies

July 18, 2008

REDUCING BACTERIAL RESISTANCE THROUGH BETTER ANTIBIOTIC PRESCRIPTION PRACTICES

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University

by

CHRISTINE OUMA
Bachelor of Science, Kutztown University of Pennsylvania, USA, May 2005

Director: DR. JAMES E. MAYS
Associate Professor, Department of Statistical Sciences and Operations Research

Virginia Commonwealth University
Richmond, Virginia
July 2008

Acknowledgement

First, I must thank God for making it all possible that this phase of my life would come to pass. I have been blessed with great advisors in Professors James Mays, Paul Brooks, and Amy Pakyz, who have given me all the support and affirmation of confidence that a graduate student could ever ask for. I also thank Dr. Ed Boone for accepting to be on my committee within such a short notice.

I would not have come this far had it not been for the support and encouragement I've received from my family over the years. I'd like to thank my brother Dennis Ouma for always being there for me. He was always patient to listen to anything without judgment, thereby enabling me to release "the steam" which would otherwise have left me frustrated. Finally I'd like to thank my parents back in Kenya for their demonstration of love. For their support and sacrifice, I dedicate this work to them.

Table of Contents

Page

List of Tables

List of Figures

Abstract

REDUCING BACTERIAL RESISTANCE THROUGH BETTER ANTIBIOTIC PRESCRIPTION PRACTICES

By

Christine Ouma,

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Mathematical Sciences with a concentration in Statistics at Virginia Commonwealth University.

Virginia Commonwealth University, 2008

Major Director: Dr. James E. Mays
Associate Professor, Department of Statistics and Operations Research

The objective of this study was to find a regression procedure that can better explain the relationship between patterns of antibiotic use and proportions of bacterial resistance. The sample for the study is comprised of 44 University Health System Consortium (UHC) member hospitals, and the data for antibiotic use and proportions of resistance are from the years 2002 to 2005. The hospitals are spread across the Northeast, South, Southwest, Midwest, and Northwest regions of the USA. Based on statistical analysis, MRSA continues to have the highest proportion of resistance among the bacteria

examined and has increased significantly since 2002. The antibiotic use in the study was measured in indices called diversity indices. There were six such measures in the study. The study, first using ordinary least squares regression, did not find one single diversity index that adequately predicted the proportion of resistance. There were also concerns that the diversity indices could be measuring the same thing, and therefore all should not be used in the model. The correlations between the three general diversity indices were strong, positive, and linear. Likewise, the three Gram-negative indices were also positively correlated with one another. Multicollinearity diagnostics also showed that there were serious dependencies among general diversity indices. Given the multicollinearity results and the correlation coefficients for the indices, it can be concluded that all six indices should not be in the same model together. Logistic regression and weighted least squares regression using the logit transformation were also performed, and just like the ordinary least squares results, there was no one single diversity index or a combination of diversity indices that adequately predicted the proportion of resistance.

Chapter 1: Introduction

1.1 Background

Bacteria are microorganisms that exist everywhere — from the great outdoors to the cleanest of homes. When bacteria get into the body, they can cause illnesses such as ear infections, strep throat, food poisoning and pneumonia. The body's immune system uses specially designed cells to locate and shut down microscopic invaders like bacteria, usually stopping them before they can cause trouble (Greenwood, 2007). People get sick — what is called a bacterial infection — when the bacteria in the body reproduce faster than the immune system can kill them. Antibiotics are used to treat bacterial infections.

Antibiotics are powerful bacteria killing drugs that help the body fight bacterial infection. Today, there are hundreds of antibiotics in use, most tailored to treat a specific kind of bacterial infection (Wax, 2008). There are antibiotics meant to kill Gram-negative bacteria and those that are meant to kill Gram-positive bacteria. Gram-negative bacteria are those bacteria that do not retain crystal violet dye in the Gram staining protocol. Gram-positive bacteria will retain the crystal violet dye when washed in a decolorizing solution. Examples of common Gram-positive bacteria are *Staphylococcus aureus* and *Streptococcus cremoris* (Greenwood, 2007). Gram-negative bacteria include a multitude of species like *Klebsiella pneumoniae, Pseudomonas aeruginosa, Escherichia coli, Enterobacter cloacae,* and *Salmonella typhi* (Greenwood, 2007).

When the usual antibiotic drugs do not seem to kill the bacteria, such bacteria are said to be resistant (Wax, 2008). Bacterial resistance makes an infection much harder to treat. Higher doses or stronger drugs may be required. In extreme cases, bacterial resistance can be fatal. According to the Centers for Disease Control and Prevention (CDC), over prescription and misuse of antibiotic drugs are the main causes of bacterial resistance. The CDC says that up to half of the roughly 100 million prescriptions for antibiotics written each year are unnecessary. In hospitals, resistance of a given bacteria is determined in the laboratory through an antibiogram. An antibiogram is the result of a laboratory testing for the sensitivity of an isolated bacterial strain to different antibiotics (Estridge, 2008). The proportion of resistant bacteria or isolates is then determined directly from the antibiogram by subtracting the percentage of isolates susceptible to the antibiotic of interest from one hundred percent and then dividing by one hundred.

Studies that have been done to examine the link between antibiotic use and resistance have traditionally relied upon the proportion of resistant isolates as the outcome of interest (Powell, 2007). A recent study conducted in an academic health-system in Brazil found that risk factors for acquiring cephalosporin- or carbapenem-resistant *P. aeruginosa* included exposure to several different antibiotics (Fortaleza et al., 2006). Other studies have shown other antibiotics to be predictive of infection with fluoroquinolone resistance in *P. aeruginosa* (Hsu et al., 2005), multi-drug resistant *P. aeruginosa* (Defez et al., 2004), imipenem-resistant *A. baumannii* (Lee et al., 2004) and methicillin-resistant *Staphylococcus aureus* (MRSA) (MacDougal et al., 2005).

Among the differing stewardship strategies being examined is the concept of antimicrobial diversity (also referred to as "heterogeneity" or "mixing"). This strategy assumes that bacterial resistance is an evolutionary response to the selective pressure from antimicrobial exposure (Powell, 2007). Diversity of antibacterial drug use is assumed to result in a mix of selective pressures on the population of bacteria such that no one pathogen is afforded a survival advantage. A recent survey of 448 hospitals found that a lack of diversity of antimicrobial use was significantly associated with higher rates of antimicrobial resistance (Zillich, 2006).

Several methods for quantifying diversity in ecological systems have been developed, including Simpson's Index of Diversity (Simpson, 1949), the Shannon-Weiner Index of Diversity (Krebs, 1989), and the Antimicrobial Homogeneity Index (Sandiumenge, 2006).

(a) Simpson's Index of diversity (D) (Simpson, 1949) is given by

$$D = \sum (1 - d_i) \text{ and } 0 \le D \le 1.$$

When D=1, we have maximum diversity and when D=0 there is no diversity;

d_i = proportion that the i^{th} antimicrobial comprises of the total volume of antimicrobials considered.

(b) The Shannon-Weiner Index of diversity (H) (Krebs, 1989) is given by

$$H = -\sum (p_i \ln p_i)$$

where p_i = proportion that the i^{th} antimicrobial comprises of the total antimicrobials considered. H ranges from 1.5 to 3.5.

(c) The Antimicrobial Homogeneity Index (AHI) (Sandiumenge, 2006) is given by

$$AHI = 1 - \left\{ \frac{n}{2(n-1)} \right\} \sum (a_i - b_i) \ ,$$

where a_i = the hypothetical proportion of antimicrobial i under maximally heterogeneous situations,

b_i = the proportion antimicrobial i comprises of the total antimicrobials considered,

n = the total number of antimicrobials considered in the score.

AHI ranges from 0 to 1.

The ideal method for measuring diversity of antimicrobial use is unknown and diversity measures differ on their sensitivity to dominant species (antibacterials) within a population. For instance, it has been noted that Simpson's Index of diversity is more accurate in determining the contribution of more dominant species towards total diversity than the Shannon-Wiener Index (Palmer and Young, 2000). It is unclear which of these formulae are best suited for quantifying diversity in terms of antimicrobial use.

1.2 Overview of the paper

The three indices that are studied in this project are Shannon-Wiener Index (SWID), Simpson's Index of Diversity (SID), and Antimicrobial Homogeneity Index (AHI). Each of the three indices is divided into two categories. The indices derived from all the antibiotics are SID, SWID, and AHI, while those derived from antibiotics that treat only Gram-negative bacteria are SIDGN, SWIDGN and AHIGN. These six diversity indices will be the explanatory variables and the proportions of resistance will be the dependent variables. The analysis of the data was done using STATA and SAS software.

1.3 Literature Review

Measures of Bacterial Resistance

Studies examining the epidemiology of bacterial resistance rely on a number of different measures, including the proportion of resistance, the frequency at which the bacteria afflict humans, and simply the presence or absence of isolates, to quantify the bacterium in question. A recent article by Schwaber et al. (2004) has cast doubt upon the traditional methods of quantifying resistance in the context of multi-hospital epidemiologic studies. The argument that they have put forth is that using the proportion of resistance as an outcome variable may not be appropriate for epidemiologists and public health officials who are interested in determining the total burden of resistance. In order to describe relationships between resistance and antibacterial usage at this aggregated level,

researchers must go beyond the proportion of resistance and conduct analyses using a rate of resistance as the outcome of interest (Powell, 2007)

Prospective Studies

There are currently only two published studies that examine the concept of antibiotic diversity as an effective stewardship policy for reducing bacterial resistance. The first was an investigation examining four different antibiotic use strategies (patient-specific, prioritization, restriction, and mixing) in a fourteen bed medical/surgical intensive care unit from March 2000 through October 2003 (Sandiumenge, 2006). The study examined these strategies in the context of empiric therapy for ventilator-associated pneumonia (VAP), which consisted of using the antibiotics anti-pseudomonal carbapenems, anti-pseudomonal cephalosporins, and piperacillin-tazobactam. The use of antimicrobials was measured as prescribed daily dose (PDD) per one hundred Intensive Care Unit (ICU) stays. The PDD was defined as the total amount of antibiotic given (in grams) divided by the usual dose prescribed in the ICU in a normal patient. These investigators devised a novel index of diversity called the Antibiotic Homogeneity Index (AHI) from the Peterson Homogeneity Index to measure the homogeneity of antibiotic use during each of the different strategy periods. Not surprisingly, the mixing period was found to have the greatest diversity of empiric drug use (AHI = 0.95). The authors identified in their conclusions that greater antibiotic heterogeneity, either through a patient specific or a mixing strategy, appeared to be associated with lower rates of resistant bacteria than the lower rates of heterogeneous use associated with cycling strategies.

The second prospective study examining antibiotic diversity was conducted in Spanish ICUs from October 2001 to June 2002 (Martinez, 2006). Martinez and colleagues explored the differences of cycling versus mixing strategies also in the setting of empiric therapy for VAP. The antibiotics considered for use were again piperacillin-tazobactam, anti-pseudomonal cephalosporins (cefepime and ceftazidime) and anti-pseudomonal carbapenems (imipenem and meropenem), but also included ciprofloxacin as an empiric choice (whereas the Sandiumenge study only added ciprofloxacin if infection with *P. aeruginosa* was suspected). The cycling period in this investigation was analogous to the prioritization cycling period of the Sandiumenge study. The authors of this study measured antibiotic use as the "mean daily prevalence of use". This was calculated by dividing the number of patients receiving an antibiotic by the number of patients in the ICU for a given day. While it was assumed that the authors considered a more diverse strategy to be one where there was an even proportion of patients receiving each of the four empiric antibiotic choices, it is unclear if they used an index of diversity to characterize this concept. Compared to the cycling period, the mixing period was marked by a significant increase in the use of anti-pseudomonal carbapemens and piperacillin-tazobactam coupled with a significant decrease in the use of anti-pseudomonal cephalosporins. Without a traditional measure of diversity, however, it is unclear if this resulted in a greater or lesser heterogeneous use of antibiotics than in the cycling period.

Chapter 2: Methodology

This study includes computation of summary statistics for the diversity indices and proportions of resistance, linear regression, logistic regression, correlation coefficients between the different diversity measures and proportion of resistance, multicolinearity diagnostics, residual analysis, and model selection for predicting the proportion of bacterial resistance.

2.1 Introduction of Linear Regression

Regression analysis refers to a collection of statistical techniques that serve as a basis for drawing inferences about relationships among variables. It is assumed that there is a functional relationship among variables and the goal is to mathematically model this relationship (Myers, 1990). For example, suppose we want to use X_1, X_2, X_3, X_4 to predict Y. The X variables are referred to as independent (or predictor or regressor) variables, and are used to determine or predict Y, called the dependent (or response variable). A suitable model is then found to explain the relationship between the X's and Y. For example, a linear regression model is $Y = \beta_0+\beta_1X_1+ \beta_2X_2+\beta_3X_3+ \beta_4X_4 + \varepsilon$, where β_0 , $\beta 1$, β_2 , β_3, and β_4 are called regression coefficients or parameters, and ε is the unexplained variability or model error.

Multiple linear regression model and assumptions

We have a dependent variable Y and *k* independent variables $X_1, X_2,\ldots,X_k$. There are *n* subjects in the experiment, and for each subject the dependent variable and each of the independent variables are measured. The multiple linear regression model is given by

$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$, where $i=1, 2, ..., n$.

This model has $p=k+1$ parameters and is linear in the parameters (Myers, 1990). The data of the experiment can be expressed as

Y	$\mathbf{X_1}$	$\mathbf{X_2}$	**…**	$\mathbf{X_k}$
y_1	x_{11}	x_{21}	…	x_{k1}
y_2	x_{12}	x_{21}	…	x_{k2}
.	.	.		.
.	.	.		.
.	.	.		.
y_n	x_{1n}	x_{2n}	…	x_{kn}

The model can then be written in matrix notation as $\mathbf{y=X\beta + \varepsilon}$

$$\mathbf{y}=\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix}, \mathbf{X}=[\mathbf{1}, \mathbf{X_1}, \mathbf{X_2},\ldots,\mathbf{X_k}]=\begin{bmatrix} 1 & x_{11} & x_{21}\ldots & x_{k1} \\ 1 & x_{12} & x_{22}\ldots & x_{k2} \\ . & . & . & . \\ . & . & . & . \\ 1 & x_{1n} & x_{2n}\ldots & x_{kn} \end{bmatrix}, \boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_k \end{bmatrix}, \text{ and } \boldsymbol{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

The **X** matrix is referred to as the model or data matrix, and $\mathbf{y=X\beta + \varepsilon}$ is the general linear regression model.

Multiple linear regression assumptions

We assume the following for the linear regression model:

- the X_{ji} are nonrandom and measured with negligible error

- $E(\varepsilon_i) = 0$, implying that the model is appropriate
- $Var(\varepsilon_i) = \sigma^2$, implying homogenous variance
- $E(\varepsilon_i\ \varepsilon_j) = 0$ for $i \neq j$, implying uncorrelated errors.

It is often assumed the the ε_i follow a normal distribution in order to make statistical inferences from the linear regression model.

Ordinary least squares

With the assumptions above, we seek an estimate **b** of the vector of parameters **β** for which

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

is minimized. In multiple regression, the predicted values for the response are

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + .. + b_k x_{ki},$$

so we wish to minimize

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - .. - b_k x_{ki})^2 .$$

In matrix notation, we want to find **b** such that (**y-Xb**)'(**y-Xb**) is minimized. To find the ordinary least squares estimates, we take partial derivatives of (**y-Xb**)'(**y-Xb**) with respect to **b,** set it equal to **0,** and solve for **b.**

$$\frac{\delta}{\delta \boldsymbol{b}} \left[(\mathbf{y-Xb})'(\mathbf{y-Xb})\right] = -2\mathbf{X}'\mathbf{y} + 2(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{0}$$

implies $(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}$, which are referred to as normal equations. **X** is an n x p matrix, with $n \geq p$. We assume that **X** is of full column rank [rank(**X**)=p], implying that the columns are

not correlated and $\mathbf{X}'\mathbf{X}$ is nonsingular. Hence $(\mathbf{X}'\mathbf{X})^{-1}$ exists, and $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ (Myers, 1990).

In this project, ordinary least squares multiple linear regression is used to find a predictive model for the proportions of resistance with diversity indices as the explanatory variables.

2.2 Introduction of Logistic Regression

Logistic Regression

Logistic Regression is a type of regression that involves binary responses (Myers, 1990). Logistic regression can be used only with two types of dependent variables:

1. A categorical response variable that has exactly two categories,
2. A continuous response variable that has values in the range 0.0 to 1.0 representing probability values or proportions.

The data structure for logistic regression is usually in either grouped or ungrouped form. In the ungrouped form, the response variable is denoted by 1 or 0. The grouped data structure usually comes from designed experiments where the experimenter can control the X's. For n different combinations of the regressor variables, we record the number r_i of successes in the n_i trials at that level, and then compute the sample proportion of successes $\hat{p}_i = \frac{r_i}{n_i}$, which is taken to be the response (Myers, 1990).

Logistic regression estimates are maximum likelihood estimates and depend on the data structure. For the i^{th} group we have n_i trials and observe r_i successes. Hence the likelihood function for the i^{th} group is

$$\binom{n_i}{r_i} P(x_i)^{r_i} \left[1 - P(x_i)\right]^{n_i - r_i} = \binom{n_i}{r_i} \left(\frac{1}{1+e^{-\mathbf{x}_i'\boldsymbol{\beta}}}\right)^{r_i} \left(1 - \frac{1}{1+e^{-\mathbf{x}_i'\boldsymbol{\beta}}}\right)^{n_i - r_i},$$

assuming the popular (nonlinear) model of

$$P(x_i) = \frac{1}{1+e^{-\mathbf{x}_i'\boldsymbol{\beta}}} + \varepsilon_i$$

(Myers, 1990). Here the errors do not have homogenous variance, so ordinary least squares is not appropriate for this model, and maximum likelihood estimates are used instead. The likelihood function for the entire sample is

$$L(\boldsymbol{\beta}, \mathbf{x}_i) = \prod_{i=1}^{n} \binom{n_i}{r_i} \left(\frac{1}{1+e^{-\mathbf{x}_i'\boldsymbol{\beta}}}\right)^{r_i} \left(1 - \frac{1}{1+e^{-\mathbf{x}_i'\boldsymbol{\beta}}}\right)^{n_i - r_i}.$$

To find maximum likelihood estimates, the natural log of the likelihood function is used, which simplifies to

$$\ln\ L(\boldsymbol{\beta}, \mathbf{x}_i) = \sum_{i=1}^{n} \left\{ \ln\binom{n_i}{r_i} + r_i \left[-\ln\left(1+e^{-\mathbf{x}_i'\boldsymbol{\beta}}\right)\right] + \left(n_i - r_i\right)\left[\ln\left(e^{-\mathbf{x}_i'\boldsymbol{\beta}}\right) - \ln\left(1+e^{-\mathbf{x}_i'\boldsymbol{\beta}}\right)\right]\right\}$$

$$= \sum_{i=1}^{n} \left\{ \ln\binom{n_i}{r_i} - r_i \ln\left(1+e^{-\mathbf{x}_i'\boldsymbol{\beta}}\right) + \left(n_i - r_i\right)\left[-\mathbf{x}_i'\boldsymbol{\beta} - \ln\left(1+e^{-\mathbf{x}_i'\boldsymbol{\beta}}\right)\right]\right\}$$

$$= \sum_{i=1}^{n} \left\{ \ln\binom{n_i}{r_i} + n_i \left[-\mathbf{x}_i'\boldsymbol{\beta} - \ln\left(1+e^{-\mathbf{x}_i'\boldsymbol{\beta}}\right)\right] - r_i\left(-\mathbf{x}_i'\boldsymbol{\beta}\right)\right\}.$$

To estimate $\boldsymbol{\beta}$, we take derivatives of $\ln\left(\boldsymbol{\beta}, \mathbf{x}_i\right)$ with respect to $\boldsymbol{\beta}$, and set them equal to $\mathbf{0}$:

$$\frac{\partial \ln L(\boldsymbol{\beta}, \boldsymbol{x}_i)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left\{ -n_i \left(\boldsymbol{x}_i - \left(\frac{e^{-\boldsymbol{x}_i' \boldsymbol{\beta}}}{1+e^{-\boldsymbol{x}_i' \boldsymbol{\beta}}} \right) \boldsymbol{x}_i \right) + r_i \boldsymbol{x}_i \right\} = \mathbf{0} .$$

Simplifying, the maximum likelihood estimates are solutions $\hat{\boldsymbol{\beta}}$ to the following:

$$\sum_{i=1}^{n} n_i \left(1 - \frac{e^{-\boldsymbol{x}_i' \boldsymbol{\beta}}}{1+e^{-\boldsymbol{x}_i' \boldsymbol{\beta}}} \right) \boldsymbol{x}_i = \sum_{i=1}^{n} r_i \boldsymbol{x}_i .$$

There are *p* (number of parameters in the model) equations with *p* unknowns, but the equations are not linear in $\boldsymbol{\beta}$ and hence cannot be solved directly. An iterative procedure is used to solve for $\boldsymbol{\beta}$*'s*.

Logit Transformation

The logit transformation is a transformation that linearizes the logistic function. Starting with the logistic function

$$P(\boldsymbol{x}_i) = \frac{1}{1+e^{-\boldsymbol{x}_i' \boldsymbol{\beta}}},$$

and solving for $\boldsymbol{x}_i' \boldsymbol{\beta}$, we find the logit transformation

$$\ln\left(\frac{P(\boldsymbol{x}_i)}{1-P(\boldsymbol{x}_i)} \right) = \boldsymbol{x}_i' \boldsymbol{\beta} .$$

Finally, $\ln\left(\frac{P(\boldsymbol{x}_i)}{1-P(\boldsymbol{x}_i)} \right)$ is regressed versus the X's using weighted least squares. In this study the response variable, which is the proportion of resistance, is transformed using the logit transformation and then regressed on the diversity indices.

2.3 Introduction of Pearson Correlation Coefficient

A correlation is a statistical measurement for the direction and strength of a linear relationship between two variables. This relationship remains linear regardless of the measurement scales (Draper and Smith, 1966). Correlation refers to the departure of two variables from independence (Miles and Shevlin, 2001). The correlation matrix is a symmetric matrix of the pairwise correlation coefficients of several variables (Frank and Todeschini, 1994).

The Pearson correlation coefficient can be obtained by dividing the covariance of the two variables by the product of their standard deviations (Myers, 1990). The values for sample correlations are known as correlation coefficients and are commonly represented by the letter "*r*". The formula for the correlation coefficient is

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}},$$

where $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{X})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{X}^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$,

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{Y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{Y}^2 = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n},$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{X})(y_i - \overline{Y}) = \sum_{i=1}^{n} x_i y_i - n\overline{X}\,\overline{Y} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}.$$

The correlation coefficient ranges between -1 and 1. A positive value for the correlation implies a positive association, that is, both variables increase or decrease together (large values of X tend to be associated with large values of Y and small values of

X tend to be associated with small values of Y). A negative value for the correlation suggests a negative or inverse association, that is, as one variable increases, the other decreases, and vice versa (large values of X tend to be associated with small values of Y and small values of X tend to be associated with large values of Y) (Edwards,1976).

If *r* is 1, it implies that all data points fall on a perfect straight line with a positive slope as shown in Figure 2.1. This is a perfect linear association (Miles and Shevlin, 2001). If *r* is -1, it implies that all data points fall on a perfect straight line with a negative slope as shown in Figure 2.2 (Miles and Shevlin, 2001).

If the correlation is 0, there is no linear relationship between the variables. An example of a plot showing a correlation of 0 is given in Figure 2.3. However, the converse is not true because the correlation coefficient detects only linear dependencies between two variables (Miles and Shevlin, 2001).

If *r* is close to 1, it indicates a strong positive fit. If *r* is close to -1, it indicates a strong negative fit. If *r* is close to 0, there is a weak linear correlation. As *r* increases from 0 to 1, this displays a moderate to strong positive relationship between the two variables. As *r* decreases from 0 to -1, this displays a moderate to strong negative relationship between the two variables. Some of the analyses in this paper also report the coefficient of determination, R^2 ($=r^2$), which measures the proportion of the variation of one variable that is predictable from the other variable.

In this project, Pearson correlation coefficients are used to measure the strength of a linear relationship between the proportion of resistance and diversity indices, and between each pair of diversity indices.

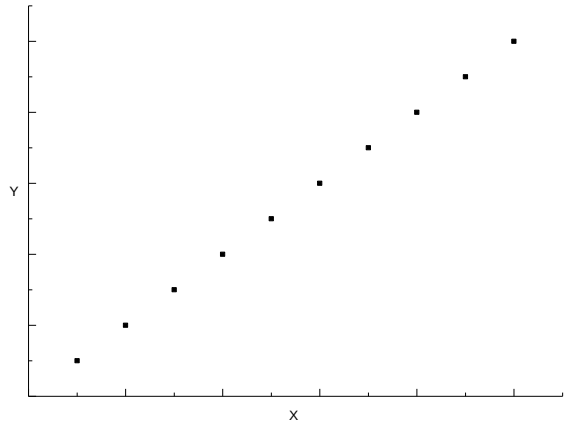

Figure 2.1: Scatterplot showing *r*=1

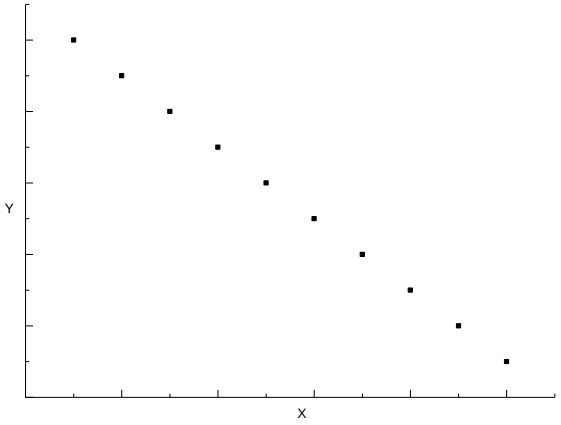

Figure 2.2: Scatterplot showing *r*=-1

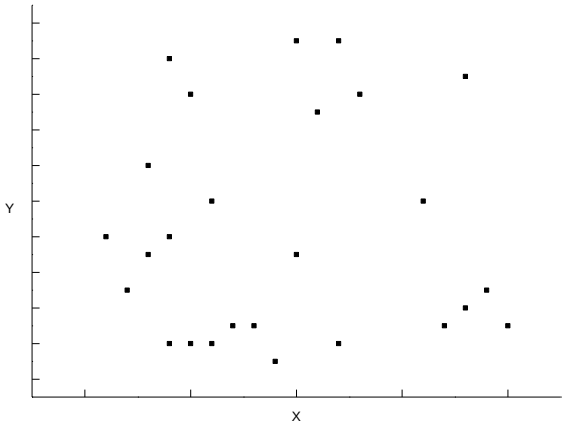

Figure 2.3: Scatterplot showing *r*=0

2.4 Introduction of Multicollinearity

Multicollinearity

Multicollinearity is a linear association among the regressor variables and exists when two or more regressors have high empirical correlations. To identify possible linear dependencies among regressors included in the model, various measures of the degree of multicollinearity are obtained. These include variance inflation factors (VIFs), Condition Index (CI), and Variance Decomposition proportions associated with the eigenvalues.

The variance inflation factor is the most common measure of multicollinearity. If R_i^2 is the coefficient of determination resulting when the predictor variable X_i is regressed on all the remaining predictor variables, the variance inflation factor for X_i (VIF_i) is given by

$$VIF_i = \frac{1}{(1-R_i^2)}.$$

The VIFs for ordinary least squares are the diagonal elements of the inverse of the simple correlation matrix. The VIFs indicate the inflation in the variance of each regression coefficient compared with a situation of orthogonality. The decision to consider a VIF to be large is essentially arbitrary. Usually, values larger than 10 suggest that multicollinearity may be causing estimation problems (Myers, 1990).

In the presence of multicollinearity, the determinant of the correlation matrix among predictor variables is very small. Because the determinant also is equal to the product of eigenvalues λ_i, the presence of one or more small eigenvalues results in a small

determinant, thereby indicating multicollinearity. A measure of multicollinearity, called the condition index (CI) is obtained for each eigenvalue by computing:

$$CI_i = \sqrt{\frac{\lambda_{\max}}{\lambda_i}}$$

where $\lambda_{\max}$ is the largest eigenvalue, and λ_i is the i^{th} eigenvalue of the correlation matrix. Large CI_i indicates dependencies among covariates because λ_i will be close to zero. Belsley (1991) suggested that a CI between 10 and 30 would indicate possible problems of multicollinearity, and a CI larger than 30 suggests the presence of multicollinearity.

Variance decomposition proportions associated with the eigenvalues indicate variables that are involved in linear dependencies, and how much of the variance of the parameter estimate is associated with each eigenvalue. Following Belsley (1991),

$$\text{Var}(\mathbf{b}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} = \hat{\sigma}^2\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}'$$

where $\hat{\sigma}^2$ is the residual variance estimate, **V** is a matrix containing the eigenvectors of **X́X**, and **Λ** is a diagonal matrix of eigenvalues of **X́X** (i.e., diag (λ_1 λ_2 . . . λ_k)). Here **X** represents the centered and scaled matrix of regressors. Writing $\mathbf{V} = v_{ij}$, the variance of the i^{th} element of **b**, the vector of regression coefficients, can be decomposed into a sum of *k* components, each associated with one eigenvalue, as follows:

$$Var(b_i) = \hat{\sigma}^2 \sum_{j=1}^{k} \frac{v^2_{ij}}{\lambda_j}$$

where *k* is the number of predictor variables. Because eigenvalues appear in the denominator, variance components associated with dependencies (small λ_j) will be relatively large compared to the other components. Thus, a high proportion of two or more

coefficients associated with the same small eigenvalue provide evidence that the corresponding dependencies are causing problems. Let

$$\tau_{ij} = \frac{\nu^2_{ij}}{\lambda_j} \text{ , and } \tau_i = \sum_{j=1}^{k} \tau_{ij} \text{ , with } i=1, ..., k.$$

Then the proportion of the variance of the i^{th} regression coefficient associated with the j^{th} component of its decomposition is obtained as follows:

$$\pi_{ji} = \frac{\tau_{ij}}{\tau_i} \text{ with } i, j = 1, ..., k.$$

2.5 Introduction of Model Selection

Variable selection is used to find the smallest, most economical, set of variables needed for estimating the dependent variable. The model selection criteria used in this study are R-square and adjusted R-square, stepwise method, forward selection, and backward elimination.

Forward selection begins by finding the variable that produces the optimum one-variable model. In the second step, the procedure finds the variable that, when added to the already chosen variable, results in the largest significant reduction in the residual sum of squares (largest increase in R^2). The third step finds the variable that, when added to the two already chosen, gives the largest significant reduction in residual sum of squares (maximum R^2). The process continues until no variable considered for addition to the model provides a reduction in sum of squares considered statistically significant at a level

specified by the user. An important feature of this method is that once a variable has been selected, it stays in the model.

Backward elimination begins by computing the regression with all independent variables included in the model. The procedure deletes from the model the variable whose coefficient has the largest insignificant *p*-value (smallest partial *F* value). The resulting equation is examined for the variable now contributing the least, which is then deleted, and so on. The procedure stops when all coefficients remaining in the model are statistically significant at a level specified by the user. With this method, once a variable has been deleted, it is deleted permanently.

The stepwise method begins like forward selection, but after a variable has been added to the model, the resulting equation is examined to see if any coefficient has a sufficiently large *p*-value to suggest that a variable should be dropped. Once all insignificant variables have been dropped, the remaining variable with the lowest significant *p*-value is added to the model and then all variables in the model are tested to see if any should be removed. This procedure continues until no additions or deletions are indicated according to the significance level chosen by the user.

2.6 Introduction of Residual Analysis

Analysis of the residuals, $e_i = y_i - \hat{y}_i$, can be used to inspect the adequacy of the model and any possible violations of assumptions. One way to do this is to inspect a plot of the residuals versus the independent variable values (x_i) or versus the predicted values ($\hat{y}_i$). If the model is appropriate, a plot of residuals versus x_i should be centered at 0 and

scattered about the line at 0. The absolute residuals should be close to 0, with what is considered “small” depending on the size of the residuals when compared to the observed *Y* data. The assumptions for ordinary least squares, which include normally distributed errors and homogenous variance, should be checked. The residuals also provide a check of whether or not the proper structure has been chosen for the model (including linear versus nonlinear).

Chapter 3: Descriptive Statistics

List of bacteria in the Study

CERKS: cephalosporin-resistant *Klebsiella* species

CERES: cephalosporin-resistant *Enterobacter* species

FQREC: fluoroquinolone-resistant *Escherichia coli*

MRSA: methicillin-resistant *Staphylococcus aureus*

CERPA: cephalosporin-resistant *Pseudomonas aeruginosa*

CPRPA: carbapenem-resistant *Pseudomonas aeruginosa*

FQRPA: fluoroquinolone-resistant *Pseudomonas aeruginosa*

PTRPA: piperacillin-tazobactam-resistant *Pseudomonas aeruginosa*

List of indices in the Study

SID: The Simpson's Index of diversity

SIDGN: The Simpson's Index of diversity Gram Negative

SWID: Shannon-Weiner Index of diversity

SWIDGN: Shannon-Weiner Index of diversity Gram Negative

AHI: Antimicrobial Homogeneity Index

AHIGN: Antimicrobial Homogeneity Index Gram Negative

3.1 Descriptive Statistics of Proportion of Resistance

The number of observations varied throughout the study period depending on the number of health-systems that agreed to participate and the number of antibiograms that were submitted during a given year. The overall mean trends of resistant bacteria in hospitals for which we could calculate proportions for all four years in the study period are depicted in Figure 3.1. Descriptive statistics (mean, median, standard deviation, maximum, and minimum) are shown in Table 3.1. As depicted in the graph, MRSA proportions of resistant isolates were the highest while CERKS proportions of resistant isolates were the lowest.

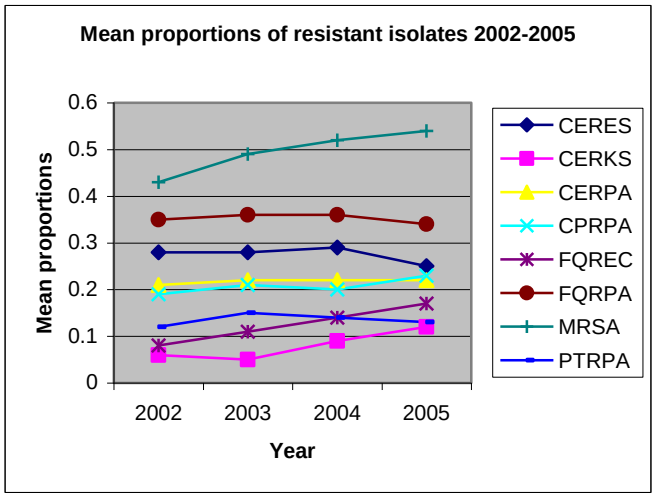

Figure 3.1 Trends in mean proportions of resistant bacteria from 2002-2005

Table 3.1: Descriptive statistics for proportions of resistant isolates from 2002-2005

Bug-Year	n	mean	median	stdev	min	max
CERES-2002	21	0.28	0.28	0.115	0.12	0.66
CERKS-2002	21	0.06	0.03	0.076	0.00	0.26
CERPA-2002	21	0.21	0.18	0.116	0.04	0.44
CPRPA-2002	27	0.19	0.18	0.084	0.04	0.42
FQREC-2002	23	0.08	0.07	0.055	0.00	0.19
FQRPA-2002	31	0.35	0.36	0.103	0.16	0.54
MRSA-2002	31	0.43	0.45	0.103	0.19	0.64
PTRPA-2002	27	0.12	0.12	0.052	0.04	0.26
CERES-2003	23	0.28	0.25	0.105	0.14	0.54
CERKS-2003	23	0.05	0.03	0.063	0.00	0.26
CERPA-2003	23	0.22	0.21	0.101	0.05	0.41
CPRPA-2003	32	0.21	0.21	0.081	0.05	0.39
FQREC-2003	25	0.11	0.09	0.062	0.02	0.28
FQRPA-2003	32	0.36	0.38	0.097	0.14	0.51
MRSA-2003	33	0.49	0.49	0.099	0.28	0.71
PTRPA-2003	30	0.15	0.14	0.066	0.03	0.29
CERES-2004	23	0.29	0.30	0.095	0.13	0.51
CERKS-2004	24	0.09	0.04	0.126	0.00	0.41
CERPA-2004	24	0.22	0.21	0.093	0.08	0.41
CPRPA-2004	31	0.20	0.20	0.099	0.05	0.48
FQREC-2004	25	0.14	0.11	0.083	0.03	0.34
FQRPA-2004	32	0.36	0.36	0.101	0.15	0.52
MRSA-2004	33	0.52	0.51	0.097	0.30	0.71
PTRPA-2004	29	0.14	0.14	0.068	0.04	0.31
CERES-2005	20	0.25	0.22	0.140	0.04	0.64
CERKS-2005	24	0.12	0.06	0.200	0.01	1.00
CERPA-2005	24	0.22	0.21	0.095	0.05	0.43
CPRPA-2005	24	0.23	0.22	0.126	0.06	0.60
FQREC-2005	24	0.17	0.15	0.096	0.00	0.39
FQRPA-2005	25	0.34	0.36	0.106	0.10	0.58
MRSA-2005	25	0.54	0.53	0.142	0.36	0.98
PTRPA-2005	25	0.13	0.11	0.072	0.01	0.33

MRSA: methicillin-resistant *Staphylococcus aureus*

The median proportion of MRSA isolates increased steadily from 2002 to 2005 as shown in Figure 3.2. The box plots show that there were three outliers in 2002, and one outlier for each of the other three years.

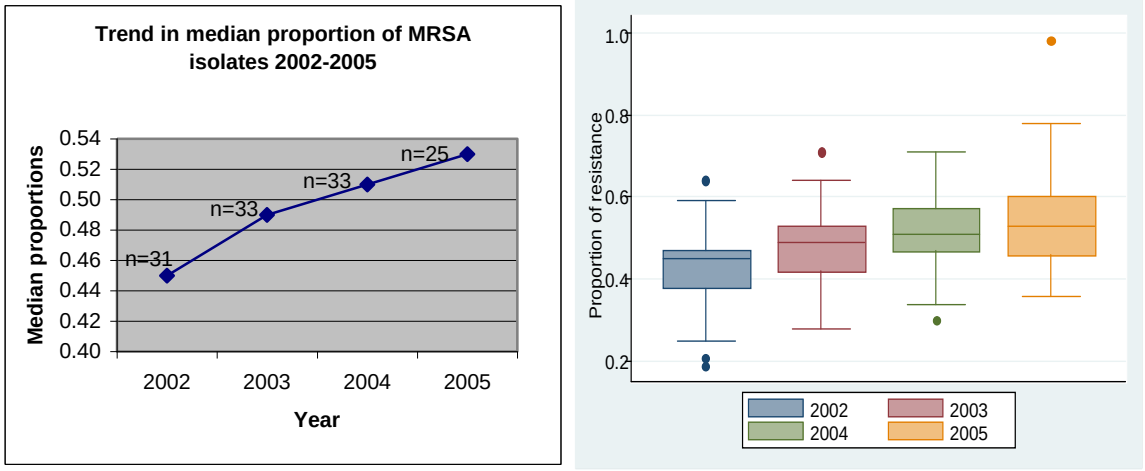

Figure 3.2 Trend in median proportion and Box plots from 2002-2005 for MRSA

CERPA: cephalosporin-resistant Pseudomonas aeruginosa

The median proportions of CERPA isolates increased steadily from 2002 to 2003 then leveled in the subsequent years as shown in Figure 3.3. The box plots show near symmetrical distributions for each of the four years that the study covered.

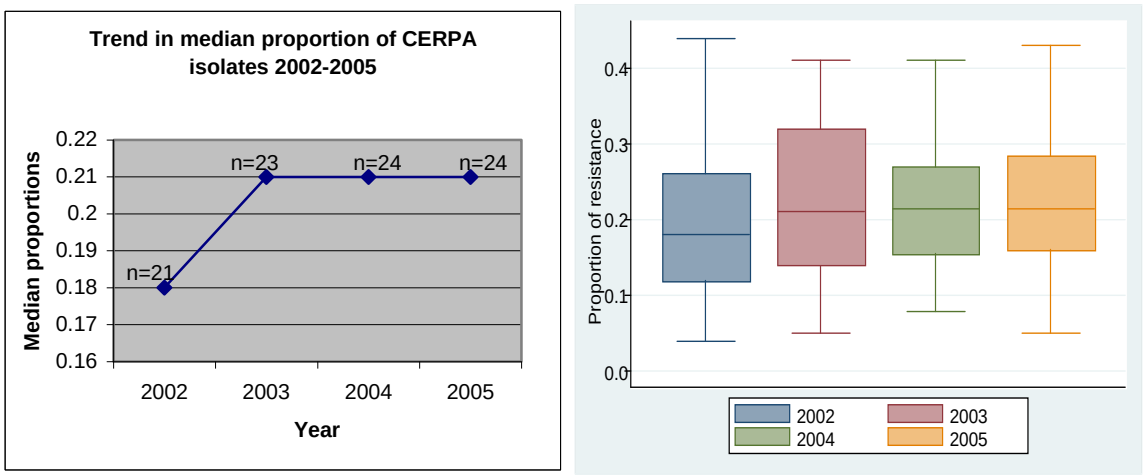

Figure 3.3. Trends in median proportions and Box plots from 2002-2005 for CERPA

CERES: cephalosporin-resistant *Enterobacter* species

The median proportion of CERES decreased from 2002 to 2003, increased in 2004, and decreased again in 2005 as shown in Figure 3.4. The box plots show an outlier observation in 2002. The variation in proportion of resistance was highest in 2005.

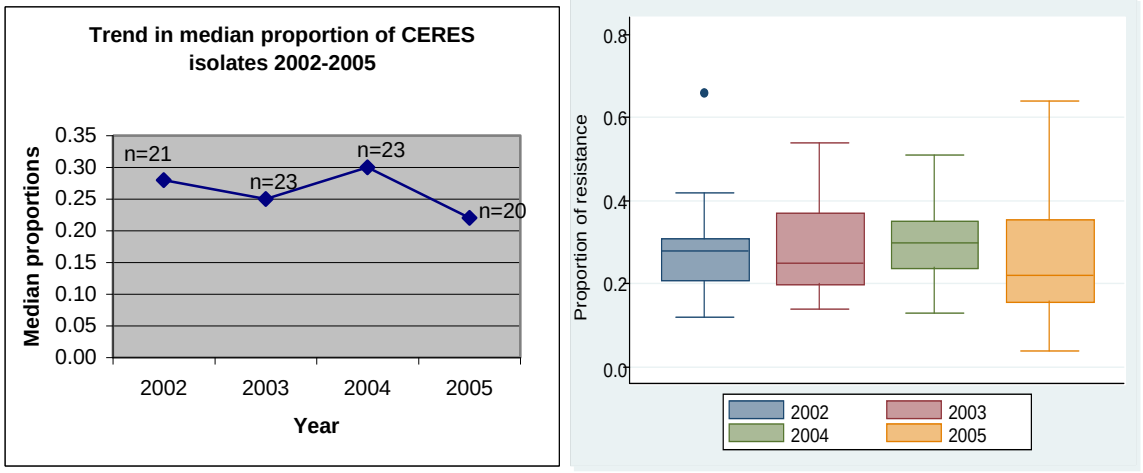

Figure 3.4. Trends in median proportions and Box plots from 2002-2005 for CERES

CERKS: cephalosporin-resistant *Klebsiella* species

The median proportion of CERKS in 2002 was similar to that of 2003 and increased in subsequent years as shown in Figure 3.5. The box plots show presence of outliers in each of the four years.

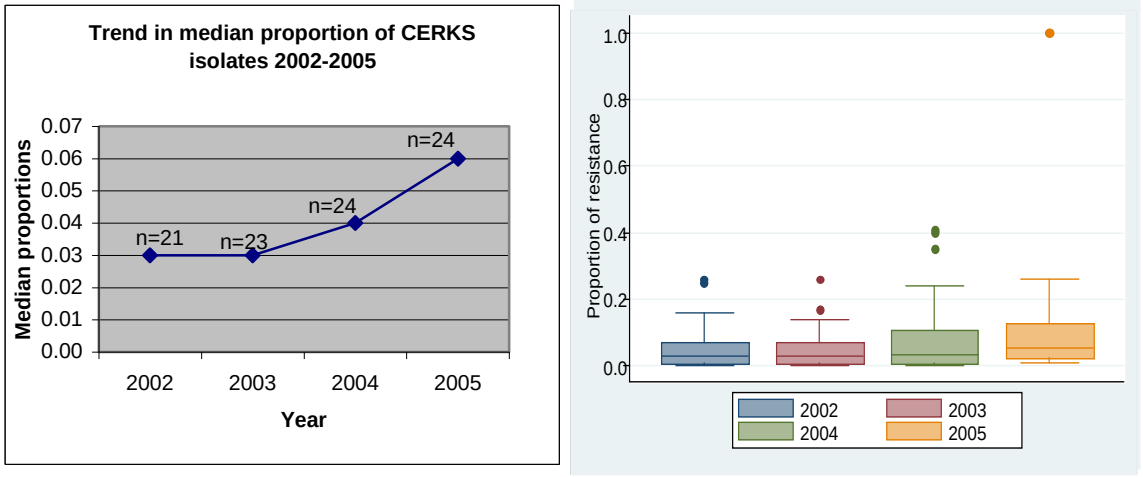

Figure 3.5. Trends in median proportions and Box plots from 2002-2005 for CERKS

CPRPA: carbapenem-resistant *Pseudomonas aeruginosa*

The median proportion for CERPA remained below 0.25 during the study period. There was a slight increase in proportion of CERPA from 2002 to 2005 as shown in Figure 3.6. Without the outliers, the distributions for 2002, 2004, and 2005 look symmetrical.

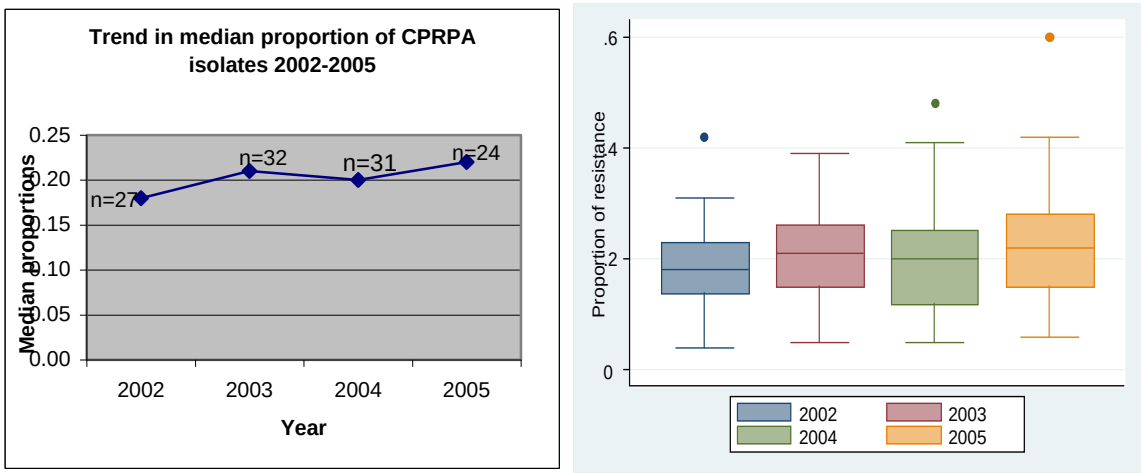

Figure 3.6. Trends in median proportions and Box plots from 2002-2005 for CPRPA

FQREC: fluoroquinolone-resistant *Escherichia coli;*

The median proportion of FQREC increased steadily from 2002 to 2005 as shown in Figure 3.7 . The box plots show presence of outliers for the years 2003 and 2004.

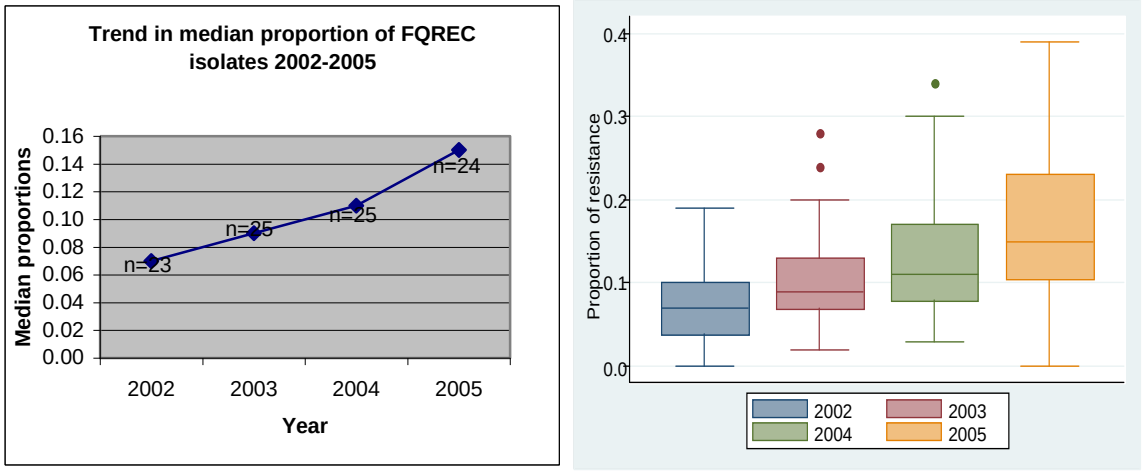

Figure 3.7. Trends in median proportions and Box plots from 2002-2005 for FQREC

FQRPA: fluoroquinolone-resistant *Pseudomonas aeruginosa*

The median proportion of FQRPA increased from 2002 to 2003, then decreased from 2003 to 2004, and finally leveled off as shown in Figure 3.8. The box plots for each of the four years show symmetrical distributions for the four years, with two outliers in 2005.

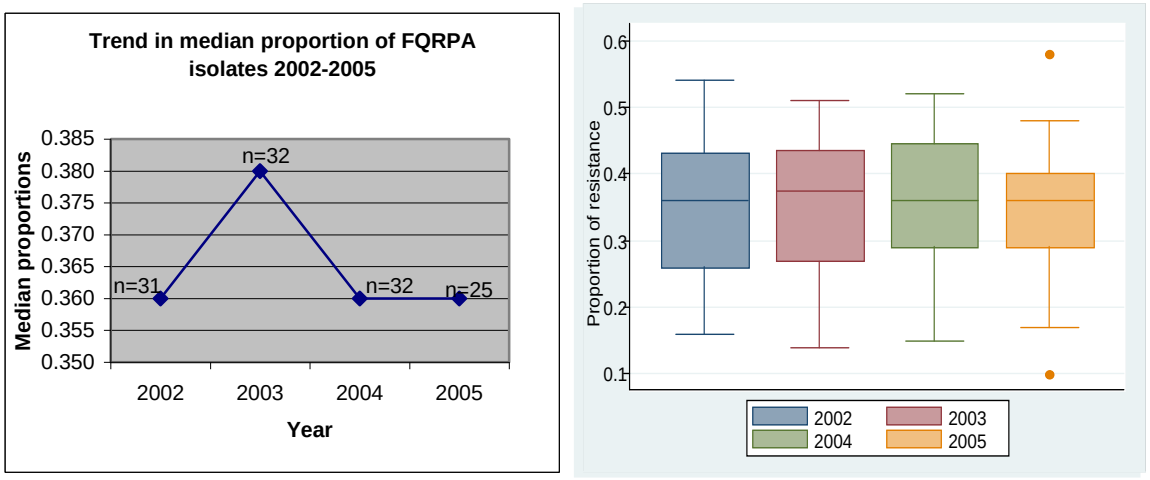

Figure 3.8. Trends in median proportions and Box plots from 2002-2005 for FQRPA

PTRPA: piperacillin-tazobactam-resistant *Pseudomonas aeruginosa:*

The median proportion of PTRPA remained low during the study period with a slight increase between 2002 and 2003, and then decreased between 2004 and 2005. The distribution for 2005 is heavily positively skewed with one outlier as shown in Figure 3.9.

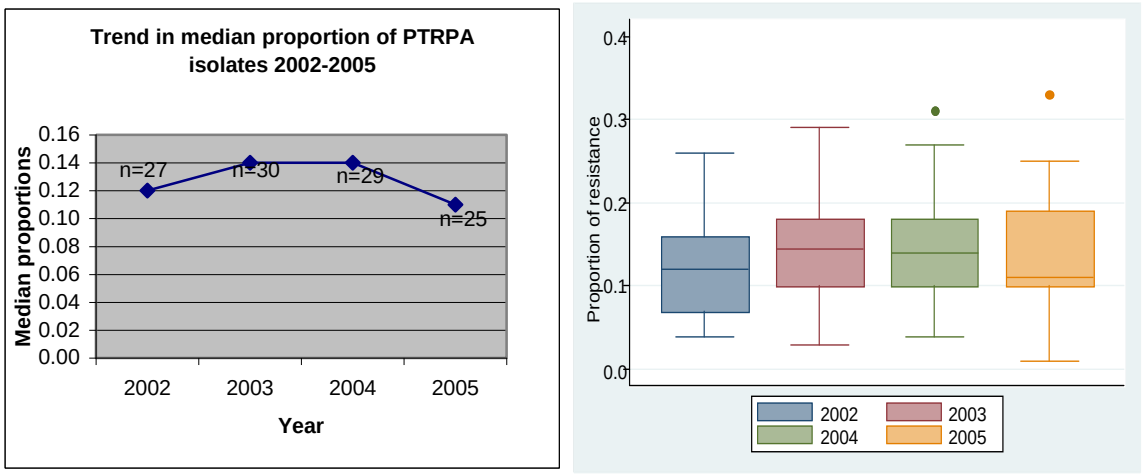

Figure 3.9 Trends in median proportions and Box plots from 2002-2005 for PTRPA

The hospitals that show up as outliers in the proportion of resistance box plots are Nevada, RW Johnson, Penn state, The Methodist Hospital, and Ohio State-East. They have proportions of resistance that are higher than the rest of the hospitals. Since most of these outliers are showing high proportions of resistance, one would expect the diversity indices associated with these outlier hospitals to be lower than the rest of the hospitals, but this is not the case. Table 3.2 shows the mean and median of diversity indices of the outlier hospitals and the mean and the median diversity indices for the rest of the hospitals.

Table 3.2: Median and mean of the diversity indices for outliers

Index-Year	Mean of outliers	Median of outliers	Mean without outliers	Median without outliers
SID-2002	0.86	0.87	0.87	0.88
SID-2003	0.86	0.87	0.88	0.88
SID-2004	0.87	0.87	0.88	0.88
SID-2005	0.88	0.87	0.88	0.88
SWID-2002	2.21	2.22	2.25	2.27
SWID-2003	2.21	2.23	2.26	2.27
SWID-2004	2.24	2.23	2.27	2.29
SWID-2005	2.26	2.25	2.26	2.28
AHI-2002	0.61	0.62	0.63	0.64
AHI-2003	0.59	0.58	0.62	0.62
AHI-2004	0.59	0.58	0.62	0.63
AHI-2005	0.61	0.59	0.64	0.65

3.2 Descriptive Statistics for the Diversity Indices

The number of diversity indices also varied throughout the study period depending on the number of health-systems that agreed to participate and number of antibiograms that were submitted during a given year. The descriptive statistics (mean, median, standard deviation, maximum, and minimum) for the six diversity indices are shown in Table 3.3. It is clear from these summary statistics that the values for each index have been very consistent across the years.

Table 3.3: Descriptive Statistics for Diversity Indices

Index-Year	n	mean	median	stdev	min	max
SID-2002	31	0.87	0.87	0.014	0.84	0.90
SID-2003	35	0.87	0.88	0.013	0.84	0.89
SID-2004	38	0.88	0.88	0.012	0.84	0.90
SID-2005	37	0.88	0.88	0.014	0.82	0.89
SIDGN-2002	31	0.59	0.60	0.071	0.39	0.68
SIDGN-2003	35	0.61	0.62	0.057	0.41	0.68
SIDGN-2004	38	0.61	0.62	0.069	0.34	0.70
SIDGN-2005	37	0.62	0.63	0.066	0.36	0.71
SWID-2002	31	2.25	2.26	0.077	2.07	2.42
SWID-2003	35	2.26	2.27	0.070	2.06	2.35
SWID-2004	38	2.26	2.27	0.068	2.07	2.38
SWID-2005	37	2.26	2.28	0.075	2.01	2.37
SWIDGN-2002	31	1.08	1.10	0.132	0.67	1.26
SWIDGN-2003	35	1.11	1.12	0.105	0.72	1.23
SWIDGN-2004	38	1.11	1.14	0.129	0.63	1.29
SWIDGN-2005	37	1.12	1.13	0.122	0.72	1.32
AHI-2002	31	0.64	0.63	0.046	0.53	0.74
AHI-2003	35	0.62	0.62	0.042	0.50	0.69
AHI-2004	38	0.62	0.63	0.044	0.49	0.68
AHI-2005	37	0.62	0.63	0.045	0.51	0.70
AHIGN-2002	31	0.61	0.62	0.091	0.39	0.75
AHIGN-2003	35	0.63	0.64	0.067	0.44	0.73
AHIGN-2004	38	0.63	0.65	0.093	0.25	0.77
AHIGN-2005	37	0.64	0.64	0.087	0.38	0.82

3.3 Box Plots for the Diversity Indices

Box plots were constructed to graphically summarize and display the data for all six diversity indices for the years 2002-2005. For SID in Figure 3.10, the datasets for the years 2002, 2003, 2004, and 2005 are all skewed to the left due an outlier present in each. The dataset for 2002 would be symmetrical without the outlier. The median value, lower quartile, and upper quartile do not vary much across the years.

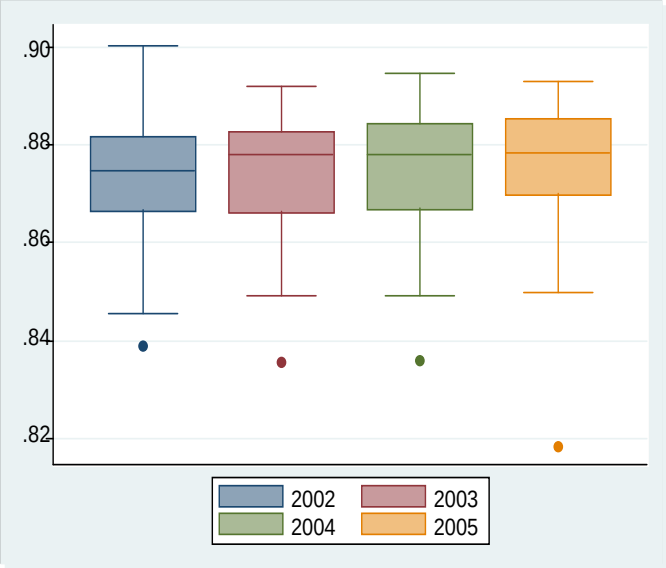

Figure 3.10: Box plots of Simpson's Index of diversity 2002-2005

SIDGN data sets for all four years are skewed to the left due to the presence of outliers as shown in Figure 3.11. Year 2005 has three outliers, 2003 and 2004 have two outliers each, and 2002 has one outlier.

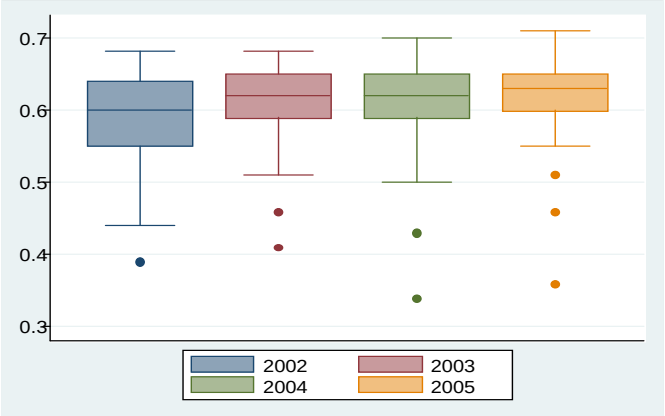

Figure 3.11 Box plots of Simpson's Index of diversity Gram Negative 2002-2005

SWID data sets for all four years are negatively skewed due to the presence of outliers, as seen in Figure 3.12.

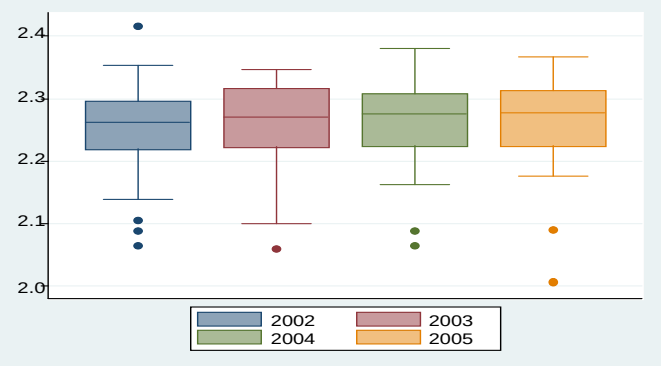

Figure 3.12: Box plots of Shannon-Weiner Index of diversity 2002-2005

SWIDGN data sets for all four years are negatively skewed due to the presence of outliers as shown in Figure 3.13.

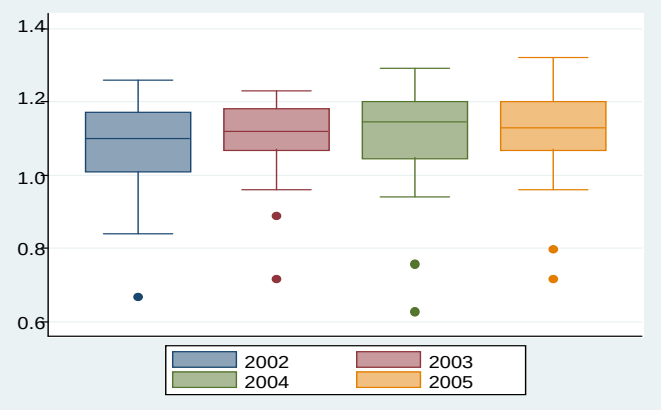

Figure 3.13: Box plots of Shannon-Weiner Index of diversity Gram Negative 2002-2005

As seen in Figure 3.14, the AHI distribution for 2002 is positively skewed and for 2005 is negatively skewed, while those for 2003 and 2004 are nearly symmetrical.

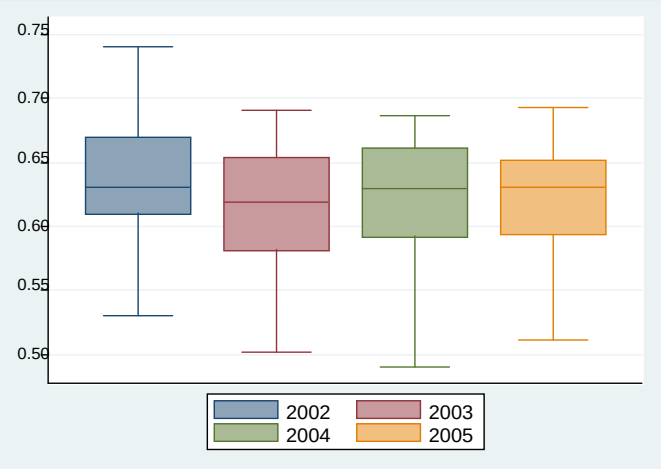

Figure 3.14: Box plots of Antimicrobial Homogeneity Index 2002-2005

The AHIGN distribution is negatively skewed for all four years, as shown on Figure 3.15.

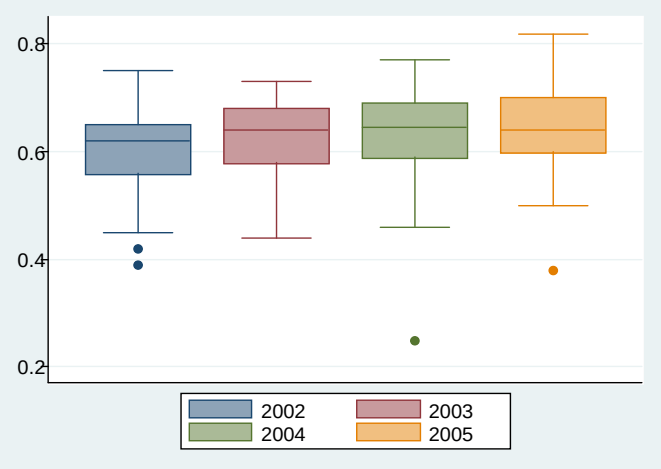

Figure 3.15: Box plots of Antimicrobial Homogeneity Index Gram Negative 2002-2005

The Hospitals that show up as outliers in the box plots for the diversity indices are Hospital 11 (SID 2003, SWID 2002-2005), Hospital 36 (SID 2002, SWID 2002), Hospital 15 (SID 2004-2005, SIDGN 2003-2005, SWID 2002 & 2004-2005, SWIDGN 2002-2003

& 2005), Hospital 6 (SIDGN 2002-2004, SWIDGN 2003-2004, AHIGN 2004), Hospital 42 (SIDGN 2005, SWIDGN 2005, AHIGN 2005), Hospital 16 (AHIGN 2005), and Hospital 38 (SIDGN 2005). Table 3.4 shows the mean and median proportions of resistance for the outlier hospitals, and mean and median proportions of resistance for the remainder hospitals for the years 2002 and 2003. Results for 2004 and 2005 are very similar to 2002 and 2003 and are not reported here. Even though most of these hospitals had unusually low diversity index values, the mean and median proportions of resistance for these hospitals are smaller than the mean and median for the rest of the hospitals. This result was unexpected.

Table 3.4: Median and mean proportions of resistance for outlier diversity indices

Bug-Year	Mean of outliers	Median of outliers	Mean without outliers	Median without outliers
CERKS-2002	0.04	0.02	0.06	0.03
CERPA-2002	0.19	0.11	0.22	0.22
CPRPA-2002	0.15	0.17	0.22	0.22
FQREC-2002	0.03	0.04	0.09	0.08
FQRPA-2002	0.30	0.28	0.36	0.39
MRSA-2002	0.39	0.44	0.43	0.46
PTRPA-2002	0.11	0.08	0.13	0.12
CERKS-2003	0.03	0.02	0.06	0.03
CERPA-2003	0.14	0.10	0.23	0.21
CPRPA-2003	0.16	0.21	0.21	0.21
FQREC-2003	0.07	0.07	0.12	0.09
FQRPA-2003	0.27	0.28	0.36	0.40
MRSA-2003	0.42	0.43	0.49	0.50
PTRPA-2003	0.09	0.08	0.15	0.15

3.4 Pearson Correlations for the Diversity Indices

Correlation coefficients for the indices were computed for the years 2002 (Table 3.5), 2003 (Table 3.6), 2004 (Table 3.7), and 2005 (Table 3.8). Indices for all drugs (SID, SWID, and AHI) have strong positive correlations between them, as do the indices for Gram-negative drugs (SIDGN, SWIDGN, AHIGN). The correlations of indices between all drugs and Gram-negative drugs range from weak (0.03 for AHI and AHIGN in 2002) to moderate (0.55 for SWID and AHIGN in 2003).

Table 3.5 Diversity indices correlations for the year 2002 (N=31)

Diversity Index	*SID*	*SIDGN*	*SWID*	*SWIDGN*	*AHI*
SID	—				
SIDGN	0.23	—			
SWID	0.97	0.25	—		
SWIDGN	0.34	0.97	0.37	—	
AHI	0.80	0.08	0.85	0.18	—
AHIGN	0.28	0.94	0.28	0.96	0.03

Table 3.6 Diversity indices correlations for the year 2003 (N=35)

Diversity Index	*SID*	*SIDGN*	*SWID*	*SWIDGN*	*AHI*
SID	—				
SIDGN	0.44	—			
SWID	0.97	0.47	—		
SWIDGN	0.51	0.97	0.54	—	
AHI	0.79	0.43	0.87	0.48	—
AHIGN	0.54	0.92	0.55	0.96	0.46

Table 3.7 Diversity indices correlations for the year 2004 (N=38)

Diversity Index	*SID*	*SIDGN*	*SWID*	*SWIDGN*	*AHI*
SID	—				
SIDGN	0.33	—			
SWID	0.97	0.36	—		
SWIDGN	0.39	0.98	0.42	—	
AHI	0.76	0.21	0.85	0.26	—
AHIGN	0.31	0.96	0.34	0.97	0.18

Table 3.8 Diversity indices correlations for the year 2005 (N=37)

Diversity Index	*SID*	*SIDGN*	*SWID*	*SWIDGN*	*AHI*
SID	—				
SIDGN	0.51	—			
SWID	0.97	0.50	—		
SWIDGN	0.52	0.98	0.53	—	
AHI	0.79	0.27	0.87	0.32	—
AHIGN	0.49	0.92	0.47	0.95	0.26

The scatterplot in Figure 3.16 for the Diversity Indices for the year 2002 shows a strong linear relationship between pairs of general indices and between pairs of Gram-negative indices. All correlations between the proportion of resistance for MRSA and the diversity indices for the year 2002 were relatively weak. Scatterplots of indices for the years 2003, 2004, and 2005, and proportions of resistance for the other seven bacteria show a similar trend as that of MRSA 2002.

The correlation coefficients between the diversity indices and the proportions of resistance of MRSA from the year 2002 to 2005 are displayed in Table 3.9. The correlation coefficients of near zero indicate almost no linear relationship between MRSA and the diversity indices, while those correlation coefficients with magnitudes between ±0.1 and ±0.36 show a weak linear relationship between MRSA and the diversity indices.

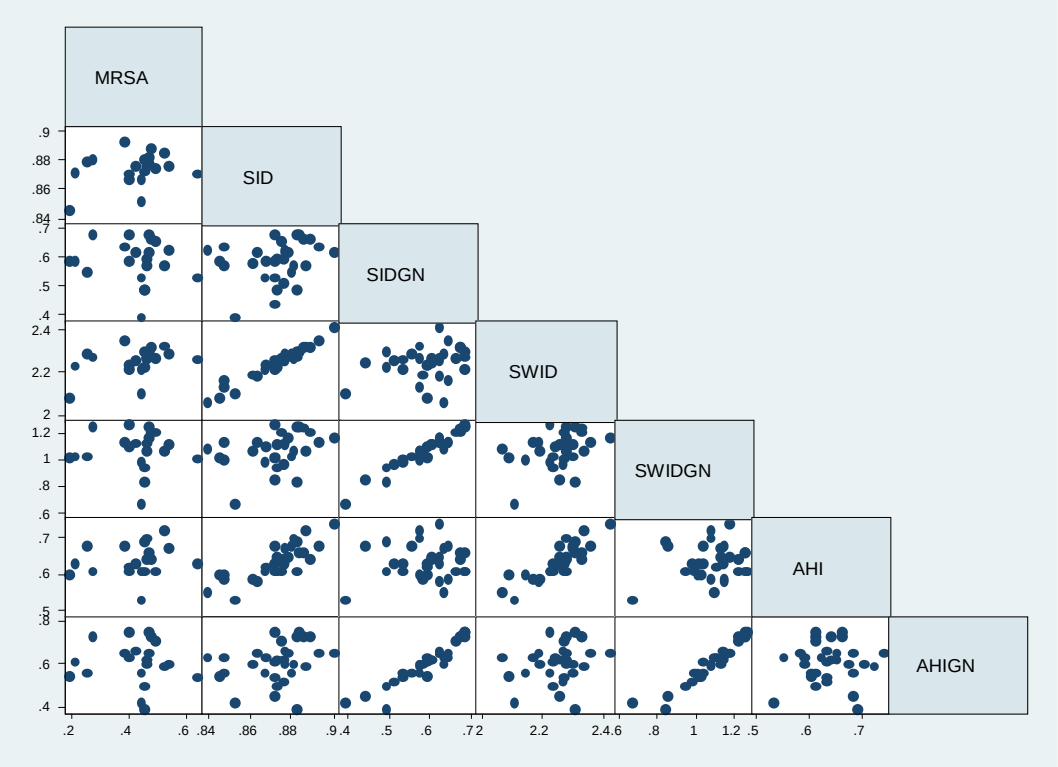

Figure 3.16: Scatterplot of Diversity Indices in 2002

Table 3.9 Correlation coefficients between Proportions of resistance for MRSA and Diversity Indices

	MRSA 2002	***MRSA 2003***	***MRSA 2004***	***MRSA 2005***
SID	0.27	0.15	0.10	-0.08
SIDGN	-0.10	0.00	0.08	0.06
SWID	0.36	0.17	0.01	-0.08
SWIDGN	-0.00	0.09	0.02	0.06
AHI	0.18	0.01	-0.23	-0.21
AHIGN	-0.04	0.06	0.03	-0.10

Chapter 4: Results and Discussion

4.1 Least Squares Multiple Regression Results

Multiple linear regression models were constructed to examine the association between measures of diversity and the proportion of resistant isolates of various bacteria within a given year. The findings of the models are presented in Table 4.1. Coefficients that are statistically significant at $\alpha=0.05$ level are bolded.

The analysis suggests that there is significant positive association (possibly due to multicollinearity) between MRSA isolates and SWIDGN in 2004, significant negative association between MRSA isolates and AHIGN in 2004, and significant negative association between MRSA isolates and AHIGN in 2005. This implies that higher levels of SWIDGN in 2004 were associated with higher proportions of MRSA, higher levels of AHIGN in 2004 were associated with lower proportions of MRSA, and finally higher levels of AHIGN in 2005 were associated with lower proportions of MRSA isolates. These results are different from the correlation results presented in Table 3.9, likely due to multicollinearity when considering numerous variables together (discussed in section 4.2).

There are sporadic significant relationships noted for diversity measures and the proportions of various resistant FQRPA isolates. In 2002 the analysis found that diversity measures SIDGN, SWIDGN, AHI, and AHIGN based on all fourteen of the antibiotic drugs and classes were significantly associated with the proportion of FQRPA isolates. In

2003 SIDGN and SWIDGN were significantly associated with FQRPA, and in 2005 SIDGN, SWIDGN, and AHI were significant in predicting the proportion of FQRPA. Conversely, in 2004 there was no association found between any of the diversity measures and the proportion of FQRPA.

The six diversity measures based on all fourteen of the antibiotic drugs and classes showed no significant associations with proportions of PTRPA, FQREC, CERES, or CERPA isolates. All the coefficients were not significant as well as the overall F-value. The only significant association between the proportion of CERKS isolates and the diversity measures was in 2004, where SID, SWID, and AHI were significant, and in 2005 where SIDGN was significant.

There were some significant relationships noted for diversity measures and the proportions of various resistant CPRPA isolates. In 2002 the analysis found that diversity measures SWID, SWIDGN, and AHI based on all fourteen of the antibiotic drugs and classes were significantly associated with the proportion of CPRPA isolates. In 2004 SIDGN and SWIDGN were significantly associated with CPRPA. There were no significant associations found between any of the diversity measures and the proportion of CPRPA isolates in 2003 and 2005.

Table 4.1. Summary table of multiple linear regression coefficients for diversity and proportion of resistant isolates by year (values in bold are significant at α=0.05).

Bacteria-Year	F-value (p-value)	SID (p-value)	SIDGN (p-value)	SWID (p-value)	SWIDGN (p-value)	AHI (p-value)	AHIGN (p-value)
	1.76	-13.10	-2.19	3.36	0.73	-0.79	0.29
MRSA-2002	(0.179)	(0.19)	(0.12)	(0.07)	(0.45)	(0.45)	(0.78)
	1.36	-8.29	-3.05	2.52	2.09	-1.57	-0.95
MRSA-2003	(0.278)	(0.28)	(0.08)	(0.16)	(0.08)	(0.15)	(0.37)
	3.12	7.56	-1.97	-0.98	**2.38**	-1.13	**-2.02**
MRSA-2004	**(0.024)**	(0.19)	(0.11)	(0.47)	**(0.01)**	(0.10)	**(0.01)**
	1.46	-0.45	-0.85	0.80	2.21	-2.20	**-2.41**
MRSA-2005	(0.262)	(0.96)	(0.73)	(0.73)	(0.14)	(0.15)	**(0.02)**
	2.76	-2.27	0.33	1.26	0.46	-0.72	-1.26
PTRPA-2002	(0.064)	(0.63)	(0.65)	(0.13)	(0.32)	(0.18)	(0.07)
	0.86	-4.06	0.22	1.36	-0.02	-1.03	0.12
PTRPA-2003	(0.539)	(0.44)	(0.85)	(0.26)	(0.97)	(0.17)	(0.87)
	0.57		0.28	-0.30	0.25	-0.40	-0.34
PTRPA-2004	(0.748)	2.98 (0.69)	(0.82)	(0.86)	(0.74)	(0.57)	(0.57)
	1.11	-2.16	-2.63	0.58	1.49	-0.30	0.01
PTRPA-2005	(0.406)	(0.69)	(0.08)	(0.67)	(0.08)	(0.72)	(0.98)
	5.12	-2.20	**-2.18**	1.58	**2.56**	**-2.35**	**-2.31**
FQRPA-2002	**(0.0056)**	(0.73)	**(0.03)**	(0.18)	**(<0.01)**	**(<0.01)**	**(<0.01)**
	3.84	-0.59	**-4.51**	1.33	**3.16**	-1.63	-1.29
FQRPA-2003	**(0.011)**	(0.92)	**(<0.01)**	(0.37)	**(<0.01)**	(0.07)	(0.15)
	2.14	6.04	-2.01	-0.17	0.88	-1.29	0.47
FQRPA-2004	(0.09)	(0.42)	(0.21)	(0.92)	(0.43)	(0.16)	(0.63)
	5.86	-4.37	**-2.83**	1.79	**1.54**	**-2.43**	0.45
FQRPA-2005	**(0.0031)**	(0.33)	**(0.03)**	(0.12)	**(0.04)**	**(<0.01)**	(0.33)
	1.55	19.09	-2.67	-2.36	3.53	-1.47	-3.58
CERPA-2002	(0.275)	(0.18)	(0.28)	(0.32)	(0.05)	(0.27)	(0.06)
	1.93	23.41	-2.32	-3.85	0.57	0.10	-0.60
CERPA-2003	(0.164)	(0.03)	(0.33)	(0.09)	(0.76)	(0.93)	(0.74)
	1.71	11.21	1.02	-2.18	0.04	1.27	-0.84
CERPA-2004	(0.182)	(0.18)	(0.59)	(0.26)	(0.97)	(0.19)	(0.29)
	1.96	-2.98	-3.12	1.53	1.84	-1.06	-0.26
CERPA-2005	(0.146)	(0.63)	(0.07)	(0.32)	(0.06)	(0.29)	(0.69)
	3.96	-7.23	-1.28	**2.57**	**1.59**	**-1.93**	-1.53
CPRPA-2002	**(0.0204)**	(0.21)	(0.16)	**(0.03)**	**(0.04)**	**(0.01)**	(0.09)
	1.61	5.99	-2.58	-0.56	1.59	-0.28	-0.59
CPRPA-2003	(0.193)	(0.34)	(0.07)	(0.70)	(0.10)	(0.74)	(0.49)
	3.86	5.82	**-4.28**	-0.67	**2.45**	-0.23	-0.38
CPRPA-2004	**(0.010)**	(0.30)	**(<0.01)**	(0.61)	**(<0.01)**	(0.74)	(0.49)
	1.09	2.44	-3.58	0.91	2.22	-2.03	-0.63
CPRPA-2005	(0.417)	(0.79)	(0.27)	(0.68)	(0.19)	(0.21)	(0.51)

	0.95	-12.30	2.16	3.63	0.52	-2.33	-2.84
CERES-2002	(0.50)	(0.47)	(0.45)	(0.22)	(0.76)	(0.17)	(0.20)
	0.69	-7.67	3.08	2.17	-0.82	-1.12	-1.28
CERES-2003	(0.66)	(0.48)	(0.25)	(0.39)	(0.65)	(0.50)	(0.41)
	0.57	-8.04	-0.68	1.95	0.96	-0.40	-1.06
CERES-2004	(0.75)	(0.41)	(0.69)	(0.36)	(0.38)	(0.68)	(0.24)
	0.89	-14.46	-2.27	4.18	1.91	-2.58	-1.14
CERES-2005	(0.54)	(0.31)	(0.45)	(0.18)	(0.29)	(0.14)	(0.38)
	1.27	-8.15	0.73	2.49	-0.29	-1.19	-0.58
CERKS-2002	(0.36)	(0.36)	(0.62)	(0.11)	(0.76)	(0.17)	(0.60)
	1.30	-9.34	-0.17	2.71	0.49	-1.82	-0.96
CERKS-2003	(0.32)	(0.16)	(0.91)	(0.08)	(0.63)	(0.06)	(0.29)
	2.82	**-24.91**	-2.70	**7.17**	2.02	**-3.49**	-1.77
CERKS-2004	**(0.04)**	**(0.02)**	(0.22)	**(<0.01)**	(0.13)	**(<0.01)**	(0.06)
	1.83	19.20	**8.69**	-4.72	-4.45	3.70	-0.47
CERKS-2005	(0.17)	(0.18)	**(0.04)**	(0.18)	(0.07)	(0.10)	(0.76)
	1.60	-7.17	0.19	1.81	0.19	-0.68	-0.57
FQREC-2002	(0.25)	(0.19)	(0.82)	(0.06)	(0.74)	(0.20)	(0.40)
	0.70	-7.33	0.13	2.08	0.43	-1.55	-0.74
FQREC-2003	(0.65)	(0.26)	(0.93)	(0.18)	(0.69)	(0.13)	(0.42)
	1.21	-5.14	-1.78	1.99	0.77	-1.56	-1.03
FQREC-2004	(0.35)	(0.50)	(0.91)	(0.27)	(0.46)	(0.09)	(0.16)
	0.74	-7.77	-1.79	2.30	1.18	-1.61	-0.30
FQREC-2005	(0.62)	(0.27)	(0.33)	(0.19)	(0.27)	(0.15)	(0.67)

4.2 Multicollinearity Results

Variance Inflation Factors, Eigenvalues, and Condition Indices

In the VIF analysis, there are numerous VIF values as shown in Table 4.2 that are greater than the cut-off point value of 10. These values indicate that there are severe multicollinearity problems within the various measures of diversity. The variance proportions and the eigenvalues indicate that there are severe dependencies throughout the study period among the measures of diversity based on all drugs (SID, SWID, and AHI). Likewise, severe dependencies exist among the measures of diversity based on Gram-negative drugs (SIDGN, SWIDGN, and AHIGN). There are no dependencies between any of the diversity indices based on Gram-negative drugs and all drugs. This means that there

are too many predictors in the model and there may be a need for at least one general index plus at least one Gram-negative index in the model.

Table 4.2. Summary table of VIFs for diversity and proportion of resistant isolates by year.

Bacteria-Year	SID	SIDGN	SWID	SWIDGN	AHI	AHIGN
MRSA-2002	16.92	12.42	16.98	33.08	2.66	25.96
MRSA-2003	26.70	12.66	39.51	28.26	4.62	12.80
MRSA-2004	27.84	16.69	43.76	39.46	4.86	17.78
MRSA-2005	26.33	26.03	39.55	35.08	4.78	8.51
PTRPA-2002	9.02	23.08	10.44	20.07	2.77	26.07
PTRPA-2003	25.40	23.39	39.51	40.64	5.24	13.10
PTRPA-2004	49.29	23.77	74.10	37.53	4.79	10.17
PTRPA-2005	30.53	18.48	36.49	23.43	3.95	14.24
FQRPA-2002	21.71	19.84	24.02	36.91	3.39	21.81
FQRPA-2003	26.90	25.41	43.37	41.84	5.79	12.49
FQRPA-2004	31.61	26.62	52.32	51.97	5.48	15.76
FQRPA-2005	26.33	26.03	39.55	35.08	4.78	8.51
CERPA-2002	41.68	18.29	32.19	42.67	4.38	41.26
CERPA-2003	33.40	9.62	46.70	24.00	5.02	12.26
CERPA-2004	34.52	35.72	55.59	46.04	5.75	9.08
CERPA-2005	24.64	25.53	36.06	34.23	4.26	8.46
CPRPA-2002	20.55	12.22	20.57	29.22	2.84	26.91
CPRPA-2003	29.29	24.43	46.72	41.61	5.83	13.11
CPRPA-2004	34.36	29.50	56.41	46.37	6.53	10.44
CPRPA-2005	26.59	37.00	39.49	44.30	5.58	7.58
CERES-2002	43.71	19.75	34.13	34.09	4.04	38.80
CERES-2003	14.13	9.85	26.29	18.40	7.12	10.60
CERES-2004	33.19	21.30	49.17	33.22	4.45	9.14
CERES-2005	31.89	25.23	22.72	36.68	3.57	37.20
CERKS-2002	19.71	134.36	11.72	51.58	5.92	58.97
CERKS-2003	39.94	9.50	53.48	19.52	4.59	11.19
CERKS-2004	30.57	16.19	42.45	23.30	4.64	8.83
CERKS-2005	40.87	25.21	50.07	42.84	5.34	16.49
FQREC-2002	37.90	31.83	31.78	42.49	4.03	30.98
FQREC-2003	32.70	32.67	53.66	62.48	7.41	14.67
FQREC-2004	34.47	30.97	54.57	44.07	5.64	8.91
FQREC-2005	21.31	32.50	26.96	37.60	3.68	7.80

Table 4.3 illustrates sample multicollinearity output for FQREC in 2002, showing large VIFs, one small eigenvalue (based on the large condition index value), and variance proportions. The three large variance proportions for SID02 (0.851), SWID02 (0.797), and AHI02 (0.798) indicate that there may be a strong linear dependency among these three measures.

Table 4.3 Sample multicollinearity diagnostics for FQREC

Variable	VIF	Eigenvalue	CI	Proportion of Variation					
				SID	SIDGN	SWID	SWIDGN	AHI	AHIGN
SID	37.90	3.994	1.000	0.001	0.001	0.001	0.001	0.005	0.001
SIDGN	31.83	1.642	1.559	0.002	0.003	0.003	0.001	0.046	0.003
SWID	31.78	0.279	3.777	0.019	0.003	0.018	0.004	0.497	0.001
SWIDGN	42.49	0.049	8.953	0.105	0.008	0.178	0.146	0.150	0.109
AHI	4.033	0.025	12.66	0.022	0.705	0.002	0.030	0.103	0.488
AHIGN	30.98	0.008	41.37	**0.851**	*0.281*	**0.797**	*0.317*	**0.798**	*0.398*

4.3 Residual Plots

If the model is appropriate, a plot of residuals versus $\hat{y}_i$ should be centered at 0 and scattered about the line at 0. The absolute residuals should be close to 0. The residual plots shown in Figure 4.1 to Figure 4.8 indicate that most, if not all, of the models are not appropriate. The pattern of scatter indicates that least squares multiple regressions may not be appropriate, and a logit transformation of the response variable or logistic regression should be used instead of linear regression.

MRSA: methicillin-resistant *Staphylococcus aureus*

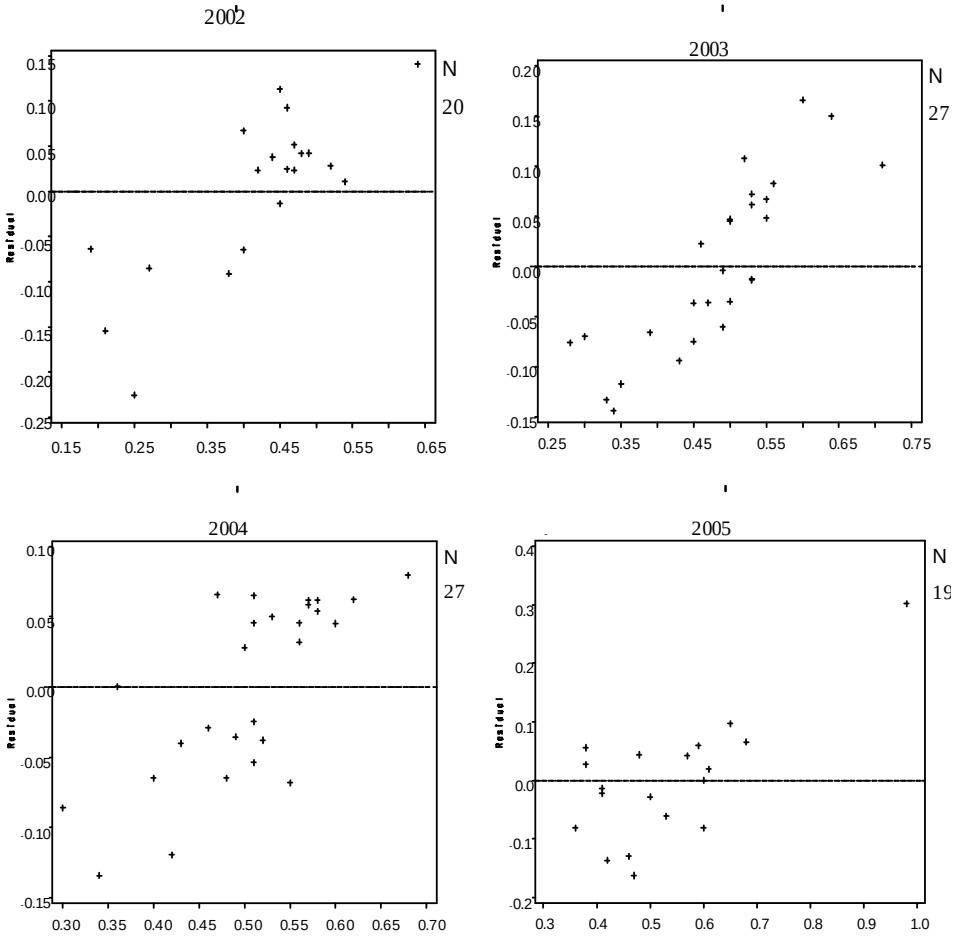

Figure 4.1 Residual plots of proportion of MRSA

CERPA: cephalosporin-resistant *Pseudomonas aeruginosa*

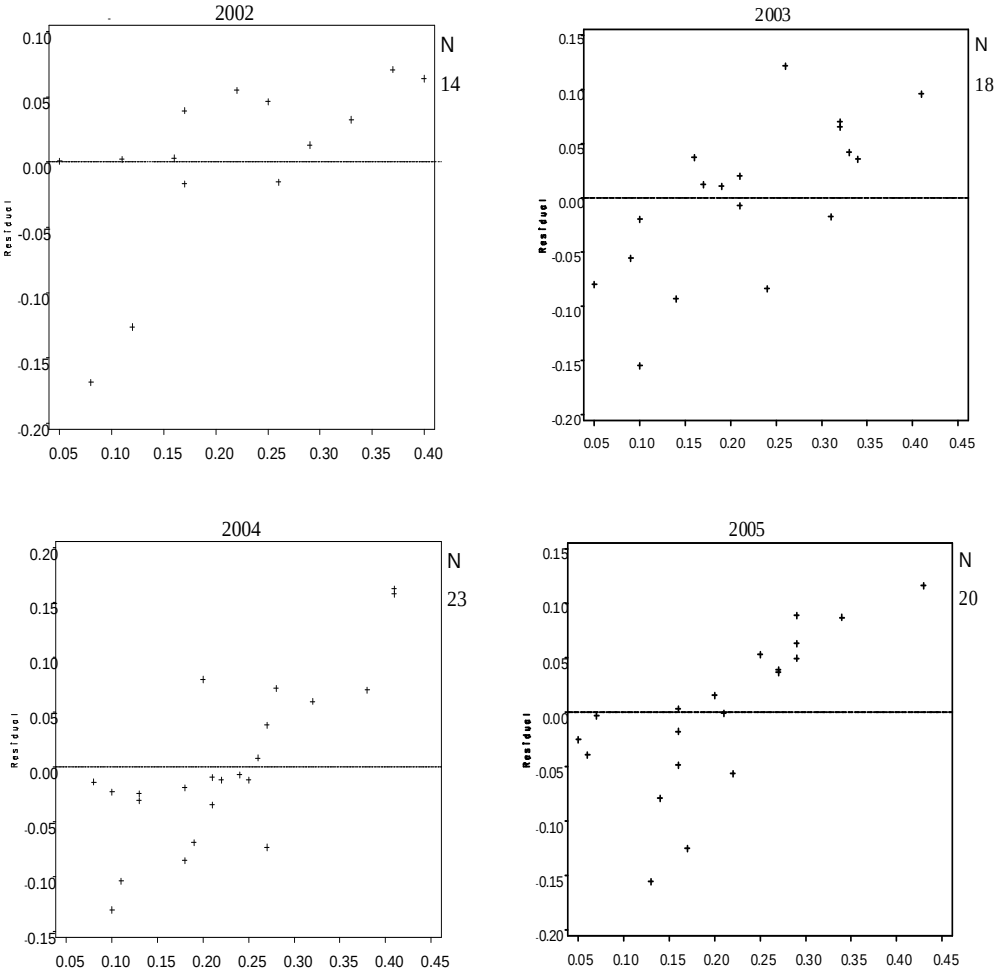

Figure 4.2 Residual plots of proportion of CERPA

CERES: cephalosporin-resistant *Enterobacter* species

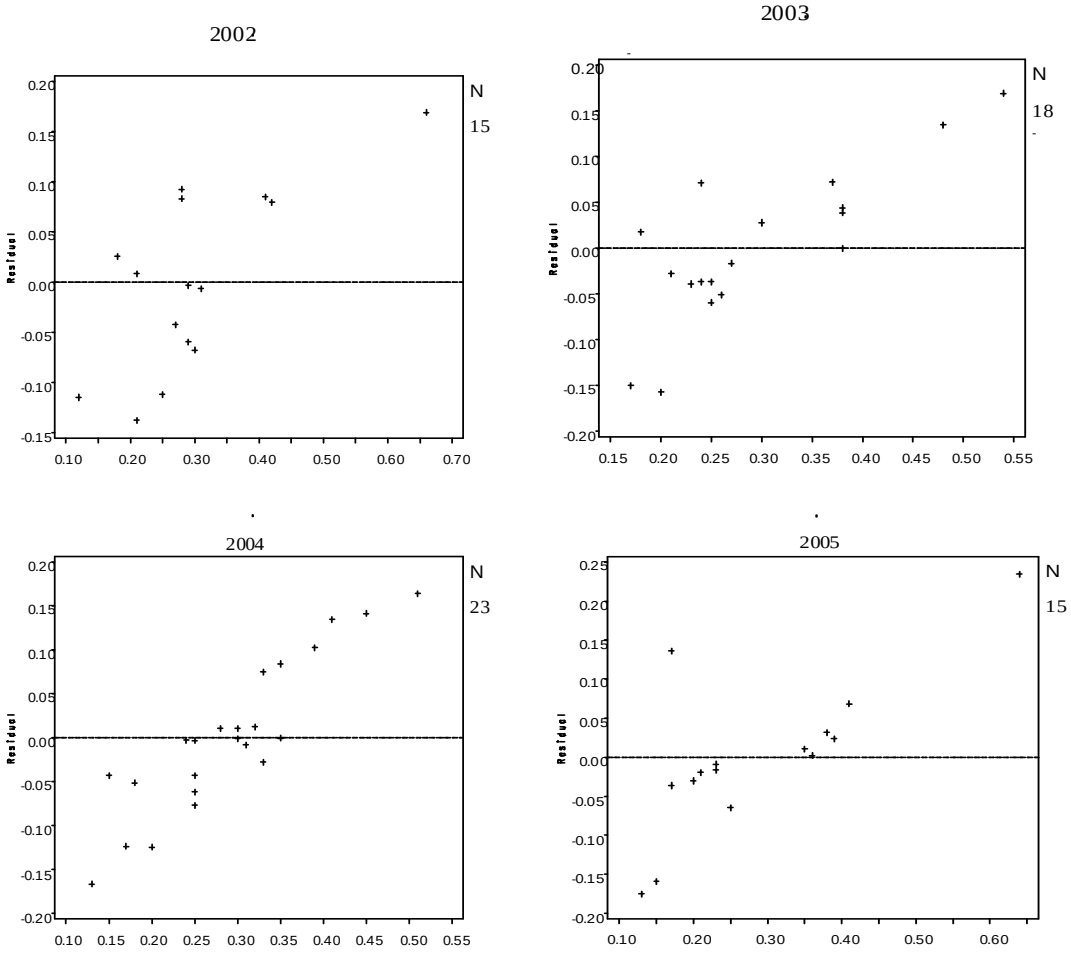

Figure 4.3 Residual plots of proportion of CERES

CERKS: cephalosporin-resistant *Klebsiella* species:

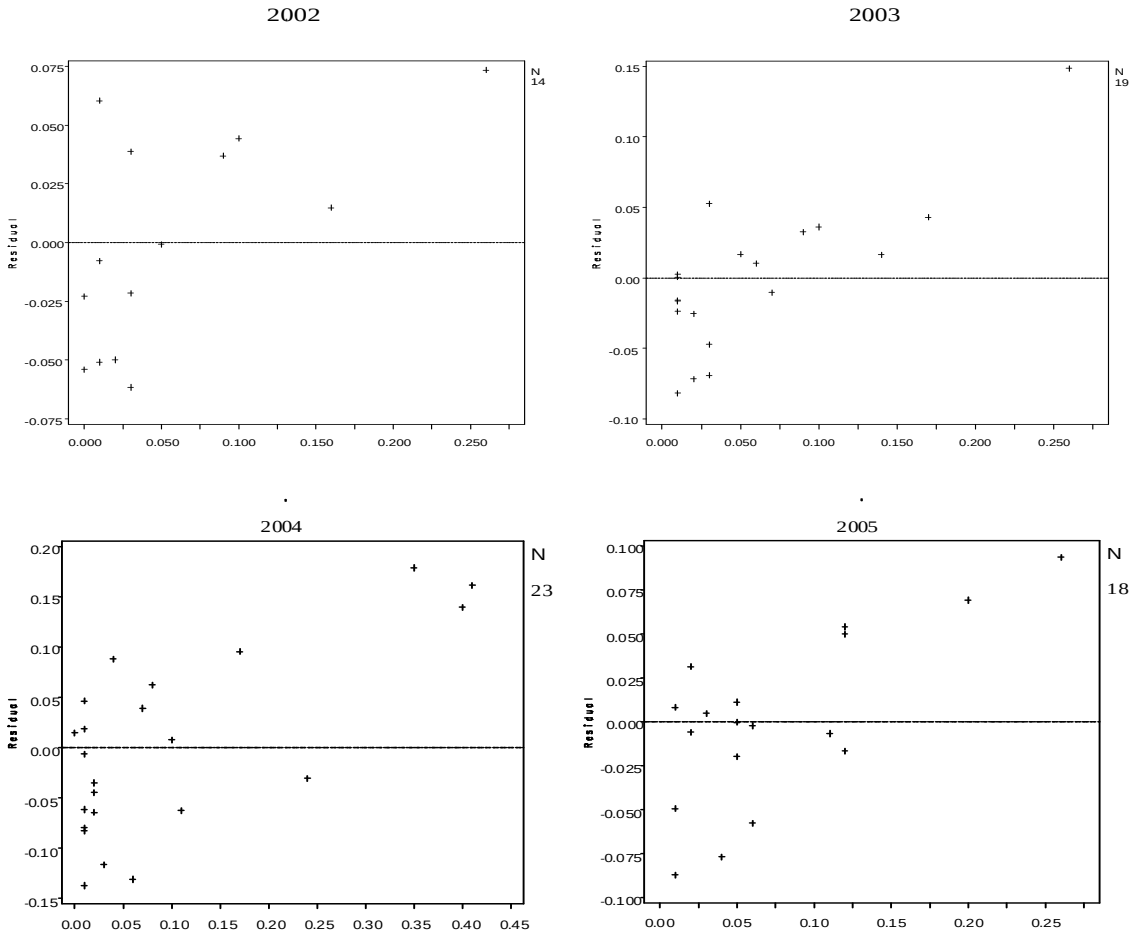

Figure 4.4 Residual plots of proportion of CERKS

CPRPA: carbapenem-resistant *Pseudomonas aeruginosa*

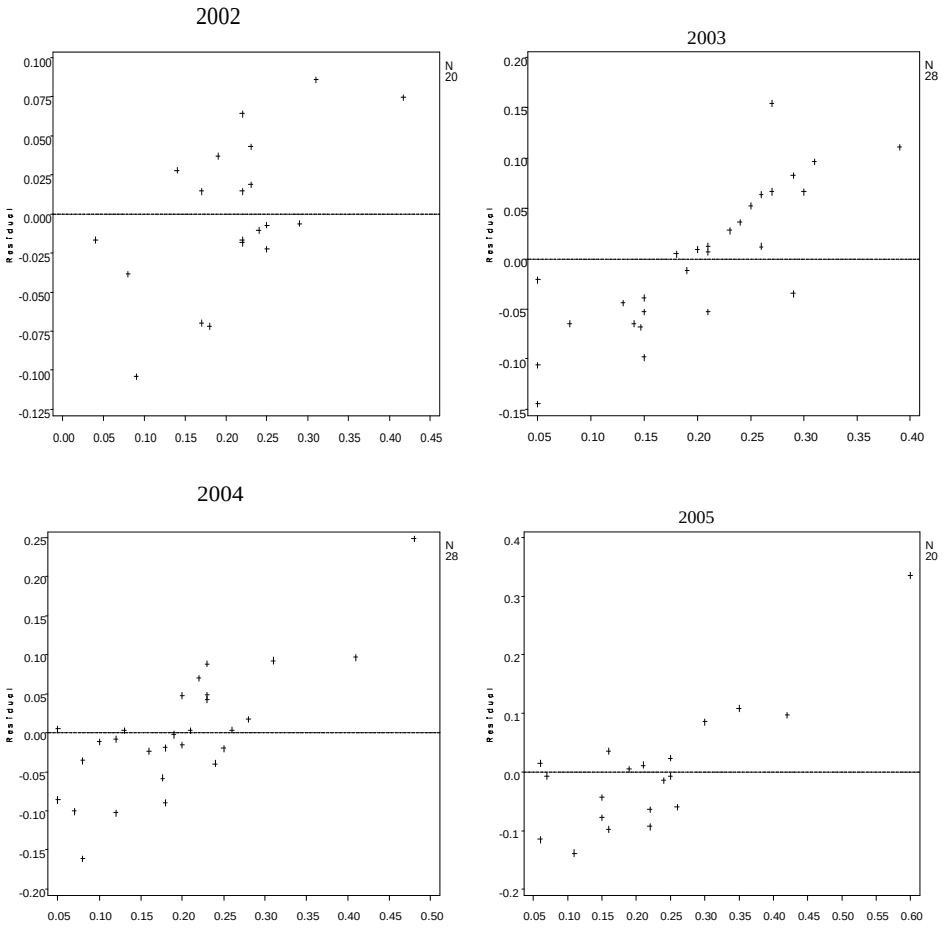

Figure 4.5 Residual plots of proportion of CPRPA

FQREC: fluoroquinolone-resistant *Escherichia coli*

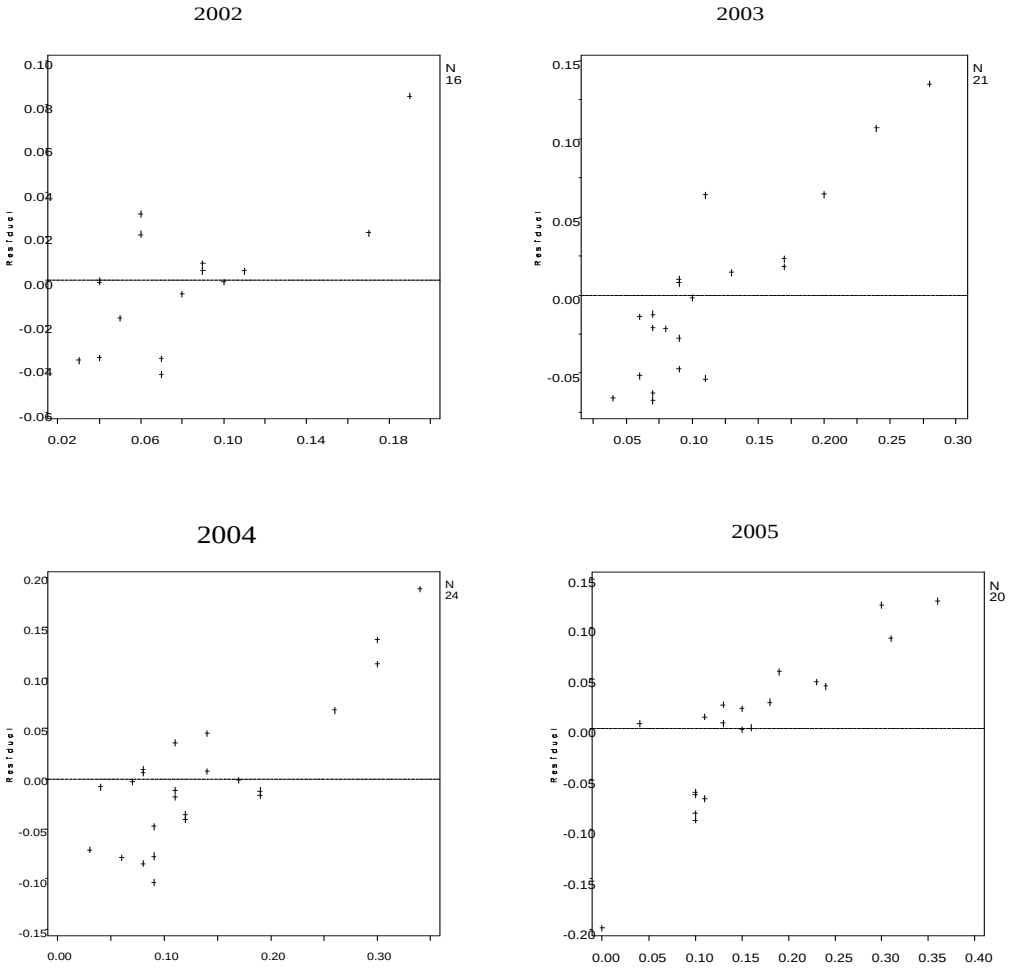

Figure 4.6 Residual plots of proportion of FQREC

FQRPA: fluoroquinolone-resistant *Pseudomonas aeruginosa*

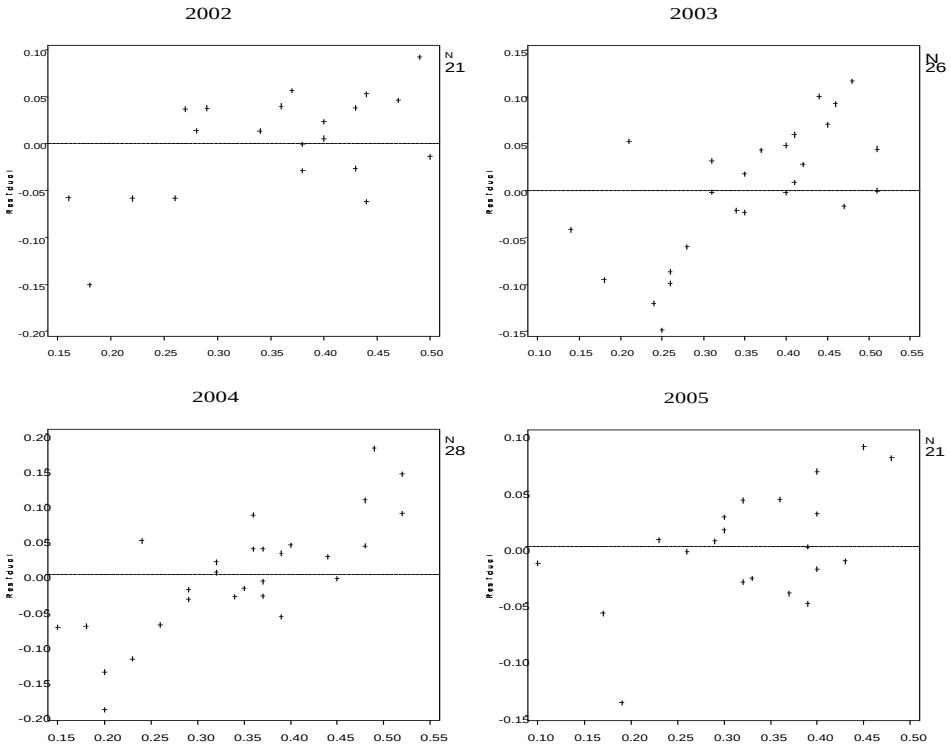

Figure 4.7 Residual plots of proportion of FQRPA

PTRPA: piperacillin-tazobactam-resistant *Pseudomonas aeruginosa:*.

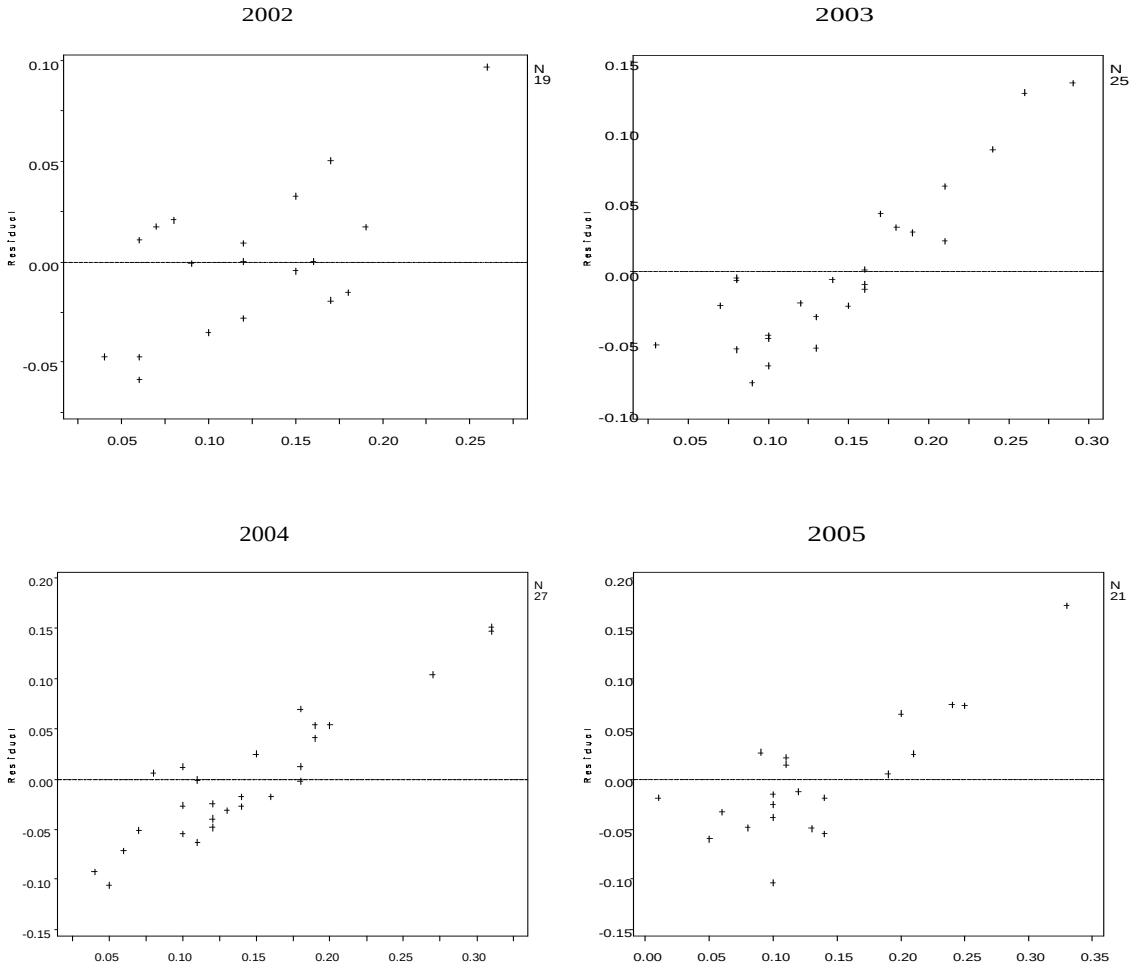

Figure 4.8 Residual plots of proportion of PTRPA

4.4 Model Selection

Using the model selection criteria of R-square, stepwise selection, backward elimination, and forward selection, the best model was chosen for each of the "Bacteria-year" combinations as shown in Table 4.4. Of the variables selected for entrance into the model by the variable selection routines (using the specified significance level to enter), there were no significant variables (at α=.05) remaining in the final models for PTRPA-2003, PTRPA-2004, CPRPA-2003, CERES-2003, CERES-2004, CERKS-2003, CERKS-2005, and FQREC-2004. For CPRPA-2005 and FQREC-2003, no independent variable met any of the model selection criteria for entrance into the model.

Table 4.4. Summary table for best model of multiple linear regression coefficients for diversity and proportion of resistant isolates by year.

Bacteria-Year	F-value (p-value)	SID (p-value)	SIDGN (p-value)	SWID (p-value)	SWIDGN (p-value)	AHI (p-value)	AHIGN (p-value)
	4.86	-16.69		**3.81**			
MRSA-2002	**(0.021)**	(0.06)		**(0.02)**			
	2.36		**-3.43**		**1.84**		
MRSA-2003	(0.116)		**(0.046)**		**(0.04)**		
	3.57				**1.38**	**-1.34**	**-2.01**
MRSA-2004	**(0.022)**	3.51 (0.06)			**(0.047)**	**(0.01)**	**(0.03)**
	3.90				**2.76**		**-3.41**
MRSA-2005	**(0.04)**				**(0.02)**		**(0.01)**
	5.06	-8.08		**2.15**			
PTRPA-2002	**(0.02)**	(0.07)		**(0.01)**			
	3.23						0.34
PTRPA-2003	(0.08)						(0.08)
	2.34		0.35				
PTRPA-2004	(0.14)		(0.14)				
	3.82		-2.53		**1.50**		
PTRPA-2005	**(0.04)**		(0.06)		**(0.03)**		
	6.50		**-2.20**	**1.21**	**2.63**	**-2.38**	**-2.41**
FQRPA-2002	**(<0.01)**		**(0.02)**	**(0.01)**	**(<0.01)**	**(<0.01)**	**(<0.01)**
	5.12		**-4.27**	**1.07**	**2.29**	-1.38	
FQRPA-2003	**(<0.01)**		**(<0.01)**	**(0.03)**	**(<0.01)**	(0.07)	
	5.18	**6.20**				**-1.38**	
FQRPA-2004	**(0.01)**	**(<0.01)**				**(0.02)**	

	8.54		**-2.93**	**0.82**	**1.92**	**-2.24**	
FQRPA-2005	**(<0.01)**		**(0.02)**	**(0.04)**	**(<0.01)**	**(<0.01)**	
	2.91	**35.12**	**-2.92**	-4.35		**-3.15**	
CERPA-2002	(0.08)	**(0.03)**	**(<0.01)**	(0.07)		**(0.04)**	
	4.80	**23.18**	**-2.00**	**-3.77**			
CERPA-2003	**(0.02)**	**(<0.01)**	**(0.02)**	**(0.01)**			
	6.57	**3.38**					
CERPA-2004	**(0.02)**	**(0.02)**					
	6.14			**0.63**			
CERPA-2005	**(0.02)**			**(0.02)**			
	4.12			**1.46**		**-1.97**	**-0.51**
CPRPA-2002	**(0.03)**			**(<0.01)**		**(0.02)**	**(0.04)**
	3.61	2.21					
CPRPA-2003	(0.07)	(0.07)					
	7.56	2.06	**-4.37**		**2.18**		
CPRPA-2004	**(<0.01)**	(0.07)	**(<0.01)**		**(<0.01)**		
	none						
CPRPA-2005		none					
	3.50	**11.74**			**-1.51**	**-4.18**	
CERES-2002	**(0.05)**	**(0.03)**			**(0.01)**	**(0.01)**	
	2.63						-0.88
CERES-2003	(0.12)						(0.12)
	1.45			0.33			
CERES-2004	(0.24)			(0.24)			
	1.66	-32.94		5.43	**4.17**		**-4.40**
CERES-2005	(0.23)	(0.06)		(0.07)	**(0.04)**		**(0.04)**
	3.82		**-1.17**	**1.90**		**-2.21**	
CERKS-2002	**(0.04)**		**(0.01)**	**(0.02)**		**(0.02)**	
	2.91		-0.90				
CERKS-2003	(0.10)		(0.10)				
	3.32	**-22.16**		**6.64**		**-3.06**	**-0.88**
CERKS-2004	**(0.03)**	**(0.03)**		**(<0.01)**		**(0.02)**	**(0.047)**
	2.97			0.35			
CERKS-2005	(0.10)			(0.10)			
	5.01			**0.58**	**0.49**	-0.59	**-1.01**
FQREC-2002	**(0.02)**			**(<0.01)**	**(0.03)**	(0.07)	**(<0.01)**
	none						
FQREC-2003							
	1.84			0.82	0.67	-1.19	-0.95
FQREC-2004	(0.16)			(0.08)	(0.16)	(0.08)	(0.17)
	2.14	**-16.41**		**3.76**		-1.65	
FQREC-2005	(0.14)	**(0.04)**		**(0.02)**		(0.07)	

One-variable model

Using model selection criteria R-square and mean square error, one best prescription practice or diversity index that could explain the proportion of resistance for

each “Bacteria-year” was selected. The model selected for each of the four years had the highest R-square and smallest mean square error. As shown in Table 4.5, there is no one single prescription practice or diversity index that fits all the bacteria. It should also be noted that none of the single variables selected were significant.

Table 4.5. Summary table for best one-variable model

Bacteria-Year	n	One-variable model: Ordinary Least squares
MRSA-2002	21	SWID
MRSA-2003	27	SWID
MRSA-2004	28	AHI
MRSA-2005	21	AHI
CERES-2002	16	SWIDGN
CERES-2003	20	AHIGN
CERES-2004	23	SWID
CERES-2005	16	SWID
CERKS-2002	16	SWIDGN
CERKS-2003	20	SIDGN
CERKS-2004	24	SWID
CERKS-2005	20	SWID
CPRPA-2002	20	SID
CPRPA-2003	28	SID
CPRPA-2004	28	SID
CPRPA-2005	20	SID
CERPA-2002	15	SIDGN
CERPA-2003	19	SID
CERPA-2004	23	SID
CERPA-2005	20	SWID
FQREC-2002	16	SWID
FQREC-2003	21	SWIDGN
FQREC-2004	24	SID
FQREC-2005	20	SWIDGN
FQRPA-2002	21	SWID
FQRPA-2003	26	SID
FQRPA-2004	28	SID
FQRPA-2005	21	AHIGN
PTRPA-2002	19	SWID
PTRPA-2003	25	AHIGN
PTRPA-2004	27	SIDGN
PTRPA-2005	20	SWIDGN

4.5 Logistic Regression Results

Since ordinary least squares regression failed to find an appropriate model that can explain the relationship between the diversity indices and the proportion of resistance, a logistic regression was performed. The data set contained the number of successes (number of bacteria that were resistant to the antibiotics) and the proportion of resistance. From this, the number of trials (number of bacteria treated with antibiotics) was arrived at by dividing the number of successes by the proportion of resistance. Also, logit transformed models were fitted using a weighted least squares regression method.

Before looking at the likelihood ratio statistics to determine whether at least one of the explanatory variables was needed in the model, a test for the appropriateness of the logistic regression was performed. Here "appropriate" means that the fitted logistic model performs as well (based on its likelihood) as a saturated model that has its number of parameters equal to the number of observations. If the test concludes that the model is "inappropriate", this does not necessarily mean that the logistic model should not be used to fit the data—it is just not as good as the saturated model. The test statistic for the appropriateness of logistic regression is called the model deviance and the hypotheses are

H_0: logistic model appropriate

H_1: logistic model inappropriate

If the model deviance is greater than $\chi^2_{\alpha,n-p}$, then the null hypothesis is rejected. For logistic regression to be appropriate, the model deviance needs to be less than $\chi^2_{\alpha,n-p}$. The association of predicted probabilities and observed responses of the proportion of

resistance will also be displayed as a percentage. The results for MRSA for each of the four years, and the results for all classes of bacteria for all four years combined are displayed below. MRSA results are given for each individual year because the proportion of resistance for MRSA was significantly different for each subsequent year. For the rest of the classes of bacteria, the years were combined because the proportion of resistance did not differ significantly from 2002 to 2005. Also given, as a measure of predictive quality of the models, are percent concordant and percent discordant values. If there are two observations that have different responses [one 'not resistant' (denoted as '0') and the other resistant (denoted as '1')], then this pair of observations is concordant if the observation with the lower actual response value gives a lower predicted response as well. They are discordant if the one with the lower actual response value gives a higher predicted response. In the results below, these values are presented as what "percent of predicted probabilities correctly fit/match the observed responses" and what percent do not. All other observations had tied predicted responses (neither concordant or discordant).

Logistic Regression for MRSA 2002

H_0: logistic model appropriate vs. H_1: logistic model inappropriate

Model deviance = 590.32

$$\chi^2_{\alpha,n-p} = \chi^2_{0.05,20-7} = 22.36$$

Since the model deviance is greater than the Chi-square critical value, the logistic regression model does not perform as well as the saturated model. The global likelihood ratio statistic is significant, which implies that at least one of the diversity indices is

important in predicting the proportion of bacterial resistance in 2002. Also, all coefficients are significant with p-values that are less than 0.001. The fitted model is

$$\hat{y} = \frac{1}{1+e^{-(13.20-44.78SID-11.11SIDGN+13.05SWID+3.58SWIDGN-3.08AHI+1.51AHIGN)}}.$$

For MRSA 2002, 52.5% of predicted probabilities correctly fit/match the observed responses while 41.5% of predicted probabilities do not fit/match the observed responses.

Logit transformed model for MRSA 2002

The overall global F-statistic is not significant with a p-value of 0.19. None of the coefficient estimates is significant at significance level α=0.05. However, if the significance level were increased to 0.1, then SIDGN and SWID both would be significant because they have p-values of 0.07. The fitted model is

$$\hat{y} = \frac{1}{1+e^{-(13.58-45.44SID-10.94SIDGN+13.07SWID+3.36SWIDGN-2.86AHI+1.70AHIGN)}}.$$

Logistic Regression for MRSA 2003

H_0: logistic model appropriate vs. H_1: logistic model inappropriate

Model deviance = 729.97

$$\chi^2_{\alpha,n-p} = \chi^2_{0.05,25-7} = 28.87$$

Since the model deviance is greater than the Chi-square critical value, the logistic regression model does not perform as well as the saturated model. The global likelihood ratio statistic is significant, which implies that at least one of the diversity indices is important in predicting the proportion of bacterial resistance in 2003. Also, all coefficients,

except the intercept, are significant with p-values less than 0.001. The fitted model is

$$\hat{y} = \frac{1}{1+e^{-(1.83-13.82SID-13.82SIDGN-18.12SWID-6.13SWIDGN+11.43AHI-3.72AHIGN)}}.$$

For MRSA 2003, 52.7% of predicted probabilities correctly fit/match the observed responses while 41.4% of predicted probabilities do not fit/match the observed responses.

Logit transformed model for MRSA 2003

The overall global F-statistic is not significant with a p-value of 0.16. However, SIDGN and SWIDGN have p-values that are less than 0.05. The fitted model is

$$\hat{y} = \frac{1}{1+e^{-(1.68-13.48SID-17.60SIDGN+6.04SWID+11.06SWIDGN-3.65AHI-4.44AHIGN)}}.$$

Logistic Regression for MRSA 2004

H_0: logistic model appropriate vs. H_1: logistic model inappropriate

Model deviance = 583.44

$$\chi^2_{\alpha,n-p} = \chi^2_{0.05,20-7} = 22.36$$

Since the model deviance is greater than the Chi-square critical value, the logistic regression model does not perform as well as the saturated model. The global likelihood ratio statistic is significant, which implies that at least one of the diversity indices is important in predicting the proportion of bacterial resistance in 2004. Also, all coefficients, except SWID, are significant with p-values that are less than 0.001. The fitted model is

$$\hat{y} = \frac{1}{1+e^{-(-12.79+21.66SID-9.54SIDGN-1.53SWID+9.46SWIDGN-4.63AHI-7.18AHIGN)}}.$$

For MRSA 2004, 52.7% of predicted probabilities correctly fit/match the observed

responses while 41.7% of predicted probabilities do not fit/match the observed responses.

Logit transformed model for MRSA 2004

The overall global F-statistic is not significant with a p-value of 0.053. However, SWIDGN and AHIGN have p-values that are less than 0.05. The fitted model is

$$\hat{y} = \frac{1}{1+e^{-(-12.38+20.64SID-9.21SIDGN-1.32SWID+9.15SWIDGN-4.61AHI-6.98AHIGN)}}$$

With a R-square of 0.43.

Logistic Regression for MRSA 2005

H_0: logistic model appropriate vs. H_1: logistic model inappropriate

Model deviance = 506.17

$$\chi^2_{\alpha,n-p} = \chi^2_{0.05,19-7} = 21.03$$

Since the model deviance is greater than the Chi-square critical value, the logistic regression model does not perform as well as the saturated model. The global likelihood ratio statistic is significant, which implies that at least one of the diversity indices is important in predicting the proportion of bacterial resistance in 2005. The coefficient estimates for SID, SIDGN, and SWID are not significant. The fitted full model is

$$\hat{y} = \frac{1}{1+e^{-(-0.24-1.49SID+0.77SIDGN+1.20SWID+2.41SWIDGN-3.87AHI-3.08AHIGN)}}.$$

For MRSA 2005, 49.0% of predicted probabilities correctly fit/match the observed responses while 41.3% of predicted probabilities do not fit/match the observed responses.

Logit transformed model for MRSA 2005

The overall global F-statistic and all the coefficient estimates are not significant. The fitted model is

$$\hat{y} = \frac{1}{1+e^{-(-0.31-0.88SID+0.84SIDGN+0.90SWID+2.25SWIDGN-3.55AHI-2.85AHIGN)}},$$

with an R-square of 0.16

Logistic Regression for MRSA, 2002-2005 combined

H_0: logistic model appropriate vs. H_1: logistic model inappropriate

Model deviance = 3084.72

$$\chi^2_{\alpha,n-p} = \chi^2_{0.05,91-7} = 106.4$$

Since the model deviance is greater than the Chi-square critical value, the logistic regression model does not perform as well as the saturated model. However, the global likelihood ratio statistic is significant, which implies that at least one of the diversity indices is important in predicting the proportion of bacterial resistance. Also, all coefficients, except the intercept, are significant with p-values that are less than 0.001. The fitted model is

$$\hat{y} = \frac{1}{1+e^{-(-0.57-8.10SID-8.60SIDGN+5.23SWID+7.07SWIDGN-6.27AHI-4.58AHIGN)}}.$$

For MRSA 2002-2005 combined, 53.4% of predicted probabilities correctly fit/match the observed responses while 44.1% of predicted probabilities do not fit/match the observed responses. Using model selection techniques, the best model for MRSA is

$$\hat{y} = \frac{1}{1+e^{-(-4.99+3.57SWID-4.41AHI-0.61AHIGN)}},$$

but the best one-variable model includes only SIDGN.

Logit transformed model for MRSA, 2002-2005 combined

The overall global F-statistic is significant with a p-value of 0.0005. The coefficient estimates for SID, SWID, and the intercept are not significant. The fitted full logit transformed model is

$$\hat{y} = \frac{1}{1+e^{-(-0.63-7.48SID-8.21SIDGN+4.90SWID+6.75SWIDGN-5.90AHI-4.37AHIGN)}}$$

with a R-square of 0.22.

The best model selected for the logit transformed model for MRSA is

$$\hat{y} = \frac{1}{1+e^{-(-3.81-8.15SIDGN+3.27SWID+6.91SWIDGN-5.55AHI-4.57AHIGN)}},$$

but the best one-variable model includes only SID.

Logistic Regression for CERES, 2002-2005 combined

The full logistic model for CERES is has a significant global likelihood ratio statistic and all the coefficient estimates are significant. The full model also has a large model deviance and therefore does not perform as well as the saturated model. The fitted model is

$$\hat{y} = \frac{1}{1+e^{-(22.99-64.81SID-2.52SIDGN+16.52SWID+4.72SWIDGN-6.88AHI-6.26AHIGN)}}.$$

This full model is also the best model when using model selection criteria. For CERES 2002-2005 combined, 56.3% of predicted probabilities correctly fit/match the observed

responses while 41.3% of predicted probabilities do not fit/match the observed responses. The best one-variable model includes only SIDGN.

Logit transformation for CERES, 2002-2005 combined

The overall F-value is significant with a p-value of 0.001. The coefficient estimates for SIDGN and SWIDGN are significant. The full fitted logit model for CERES is

$$\hat{y} = \frac{1}{1+e^{-(20.10-57.84SID-2.69SIDGN+14.94SWID+5.15SWIDGN-6.38AHI-6.73AHIGN)}}.$$

The best model selected for CERES is

$$\hat{y} = \frac{1}{1+e^{-(27.91-73.82SID+17.25SWID-5.09AHI)}},$$

with all variables in this model being significant. The best one-variable model includes only SWID.

Logistic Regression for CERPA, 2002-2005 combined

The full logistic model for CERPA has a significant global likelihood ratio statistic and all the coefficient estimates, except SWID, are significant. The full model also has a large model deviance and therefore does not perform as well as the saturated model. The full fitted logistic model for CERPA is

$$\hat{y} = \frac{1}{1+e^{-(-19.85+20.38SID-5.67SIDGN+0.93SWID+6.85SWIDGN-2.88AHI-6.03AHIGN)}}$$

For CERPA 2002-2005 combined, 56.2% of predicted probabilities correctly fit/match the observed responses while 40.8% of predicted probabilities do not fit/match the observed responses. The best selected logistic model for CERPA is

$$\hat{y} = \frac{1}{1+e^{-(-21.28+24.15SID-5.65SIDGN+6.91SWIDGN-2.54AHI-6.09AHIGN)}}$$

Logit transformed model for CERPA, 2002-2005 combined

The logit transformed full model for CERPA has a significant global F-value and none of the coefficient estimates are significant. The full model is

$$\hat{y} = \frac{1}{1+e^{-(-21.12+22.04SID-6.20SIDGN+0.67SWID+6.01SWIDGN-2.37AHI-4.44AHIGN)}}.$$

The best model selected for CERPA has only the variable SID:

$$\hat{y} = \frac{1}{1+e^{-(-20.67+22.72SID)}}.$$

Logistic Regression for CERKS, 2002-2005 combined

The full logistic model for CERKS has a significant global likelihood ratio statistic and all the coefficient estimates are significant. The full model also has a large model deviance and therefore does not perform as well as the saturated model. The full fitted model is

$$\hat{y} = \frac{1}{1+e^{-(-11.88-24.82SID+22.70SIDGN+19.10SWID-8.95SWIDGN-13.85AHI-11.36AHIGN)}}.$$

For CERKS 2002-2005 combined, 69.8% of predicted probabilities correctly fit/match the observed responses while 28.2% of predicted probabilities do not fit/match the observed responses. The best selected logistic model for CERKS is

$$\hat{y} = \frac{1}{1+e^{-(-12.41+10.85SIDGN+5.69SWID-14.88AHIGN)}},$$

and the best one-variable model includes only SIDGN.

Logit transformed model for CERKS, 2002-2005 combined

The coefficient estimates as well as the overall F-statistic for the full logit transformed model for CERKS are not significant. The full model is

$$\hat{y} = \frac{1}{1+e^{-(-5.53-32.89SID+22.20SIDGN+18.88SWID-9.62SWIDGN-10.41AHI-11.05AHIGN)}}.$$

The best selected logit transformed model for CERKS has AHIGN and SWID variables, which are all significant:

$$\hat{y} = \frac{1}{1+e^{-(-11.88+6.41SWID-8.66AHIGN)}},$$

and the best one-variable model includes only AHIGN.

Logistic Regression for CPRPA, 2002-2005 combined

The full logistic model for CPRPA has a significant global likelihood ratio statistic and all the coefficient estimates are significant. The full model also has a large model deviance and therefore does not perform as well as the saturated model. The full fitted model is

$$\hat{y} = \frac{1}{1+e^{-(-11.7-1.47SID-15.85SIDGN+7.73SWID+5.96SWIDGN-6.45AHI+1.96AHIGN)}}.$$

For CPRPA 2002-2005 combined, 57.6% of predicted probabilities correctly fit/match the observed responses while 40.1% of predicted probabilities do not fit/match the observed responses. The best selected logistic model for CPRPA is

$$\hat{y} = \frac{1}{1+e^{-(-12.34-15.81SIDGN+7.41SWID+5.96SWIDGN-6.36AHI+1.94AHIGN)}},$$

and the best one-variable model includes only AHIGN.

Logit transformed model for CERPA, 2002-2005 combined

The full logit transformed model for CPRPA has an overall significant F-value with insignificant coefficient estimates for SID and AHIGN. The full logit transformed model is

$$\hat{y} = \frac{1}{1+e^{-(-8.87-7.04SID-14.22IDGN+8.56SWID+5.11SWIDGN-6.28AHI+2.08AHIGN)}}.$$

The best selected logit transformed model for CPRPA is

$$\hat{y} = \frac{1}{1+e^{-(-11.76-13.54SIDGN+6.91SWID+6.21SWIDGN-6.10AHI)}},$$

where all the variables are significant with p-values less than 0.01. The best one-variable model includes only SID.

Logistic Regression for FQREC, 2002-2005 combined

The full logistic model for FQREC has a significant global likelihood ratio statistic and all the coefficient estimates are significant. The full model also has a large model deviance and therefore does not perform as well as the saturated model. The full fitted model is

$$\hat{y} = \frac{1}{1+e^{-(5.01-32.56SID-2.53SIDGN+11.57SWID+4.36SWIDGN-8.21AHI-4.72AHIGN)}}.$$

For FQREC 2002-2005 combined, 55.4% of predicted probabilities correctly fit/match the observed responses while 40.9% of predicted probabilities do not fit/match the observed responses. The best selected logistic model for FQREC is

$$\hat{y} = \frac{1}{1+e^{-(7.49-38.13SID+12.43SWID-6.97AHI)}},$$

and the best one-variable model includes only AHI.

Logit transformed model for FQREC, 2002-2005 combined

The full logit transformed model for FQREC has an overall significant F-value with significant coefficient estimates for SWID and AHI. The full logit transformed model is

$$\hat{y} = \frac{1}{1+e^{-(5.98-34.72SID-2.10SIDGN+11.93SWID+3.71SWIDGN-8.44AHI-3.45AHIGN)}}.$$

The best logit transformed model for FQREC is

$$\hat{y} = \frac{1}{1+e^{-(9.57-43.64SID+13.85SWID-7.57AHI)}},$$

and the best one-variable model includes only SWID.

Logistic Regression for FQRPA, 2002-2005 combined

The full logistic model for FQRPA has a significant global likelihood ratio statistic and all the coefficient estimates are significant. The full model also has a large model deviance and therefore does not perform as well as the saturated model. The full fitted model is

$$\hat{y} = \frac{1}{1+e^{-(-3.05-9.32SID-15.08SIDGN+6.80SWID+9.74SWIDGN-7.26AHI-2.87AHIGN)}}.$$

For FQRPA 2002-2005 combined, 56.1% of predicted probabilities correctly fit/match the observed responses while 41.5% of predicted probabilities do not fit/match the observed responses. The best selected logistic model for FQRPA is

$$\hat{y} = \frac{1}{1+e^{-(-1.34-15.33SID+7.79SWID-5.50AHI)}},$$

and the best one-variable model includes only AHI.

Logit transformed model for FQRPA, 2002-2005 combined

The full logit transformed model for FQRPA has an overall significant F-value with insignificant coefficient estimates for SID and AHIGN. The full logit transformed model is

$$\hat{y} = \frac{1}{1+e^{-(-3.20-8.33SID-14.58SIDGN+6.38SWID+9.47SWIDGN-6.91AHI-2.83AHIGN)}}.$$

The best logit transformed model for FQRPA is

$$\hat{y} = \frac{1}{1+e^{-(-7.62+4.34SWID-4.39AHI)}},$$

and the best one-variable model includes only SID.

Logistic Regression for PTRPA, 2002-2005 combined

The full logistic model for PTRPA has a significant global likelihood ratio statistic and all the coefficient estimates are significant except AHIGN. The full model also has a large model deviance and therefore does not perform as well as the saturated model. The full fitted model is

$$\hat{y} = \frac{1}{1+e^{-(9.66-42.68SID-3.30SIDGN+12.42SWID+2.59SWIDGN-4.27AHI-0.74AHIGN)}}.$$

For PTRPA 2002-2005 combined, 54.4% of predicted probabilities correctly fit/match the observed responses while 41.6% of predicted probabilities do not fit/match the observed responses. The best selected logistic model for PTRPA is

$$\hat{y} = \frac{1}{1+e^{-(9.76-44.34SID+13.15SWID-4.07AHI)}},$$

and the best one-variable model includes only AHI.

Logit transformed model for PTRPA, 2002-2005 combined

For the full logit transformed model for PTRPA below, the overall F-value is significant, only SID and SWID are significant individually:

$$\hat{y} = \frac{1}{1+e^{-(6.62-34.23SID-1.91SIDGN+10.24SWID+1.72SWIDGN-3.13AHI-0.63AHIGN)}}.$$

The best selected logit transformed model for PTRPA includes SWID and SID, and their coefficient estimates are significant:

$$\hat{y} = \frac{1}{1+e^{-(4.28-26.77SID+7.68SWID)}}.$$

The best one-variable model includes only SID.

One-variable models

The best one-variable models from logistic regression and logit transformed models are displayed in Table 4.6. All of these one-variable models are significant in predicting the proportion of resistant isolates.

Table 4.6. Logistic Regression summary table for best one-variable model

Bacteria-Year	n	Logistic Regression	Weighted Least Squares
MRSA-2002	21	AHI	SWID
MRSA-2003	27	AHI	SWID
MRSA-2004	28	AHI	AHI
MRSA-2005	21	AHI	AHI
MRSA-All	91	SIDGN	SID
CERES-All	75	SIDGN	SWID
CERKS-All	73	SIDGN	AHIGN
CPRPA-All	92	AHIGN	SID
CERPA-All	92	AHIGN	SID
FQREC-All	75	AHI	SWID
FQRPA-All	90	AHI	SID
PTRPA-All	88	AHI	SID

Chapter 5: Conclusions

MRSA continues to have the highest rates of resistance among the organisms examined and has increased significantly since 2002. Trends in other proportions of resistance appear to be stable or declining. There are situations where the overall F-value, which indicates that at least one of the prescription practices or diversity indices is important in predicting the proportion of resistance, is not significant while at least one of the partial t-tests is significant. This does not happen quite often, but one possible explanation why it happened in this data set is that there is very little variation in the data set, so the partial t-tests can be significant because with only 1 degree of freedom it is not required that much variation be explained. The ordinary least squares regression procedure did not find one single prescription practice or diversity index that fits all the bacteria for each of the four years. However, the diversity indices SWID and SID did occur more often than the other measures in these single variable models, giving possible evidence that these two indices may be the 'better' measures of diversity in this specific application.

A logistic regression model and a logit transformed model (using weighted least squares regression) were fitted for each of the four years for MRSA, and for all the years combined for the eight classes of bacteria in the study. The logistic regression results for the eight classes of bacteria in this study were consistent. That is, for a given bacteria, the coefficient estimates for the full logistic regression model and the full logit transformed model were similar. Except for CERKS, the overall F-values for the logit transformed

models for each index were significant when combining the four years together. A test for the appropriateness of the logistic model showed that the fitted logistic regression model did not perform as well as a saturated model for modeling the proportion of resistance versus the diversity indices for each of the four years and for all the data sets combined. This outcome was surprising because the global likelihood ratio statistics for the models for all the years combined showed that at least one of the diversity indices was necessary in predicting the proportion of resistance. In fact, when all the data sets for the four years were combined (and for MRSA for each year individually), there was a single significant diversity index for each type of bacteria that best explained the proportion of resistance, as displayed on Table 4.6. For logistic regression, AHI was the most commonly selected index, followed by SIDGN and AHIGN. For the logit transformation, there was a variety of indices chosen for the one-variable models.

Though this data set contains information for four years, there is considerable variation in the number of hospitals that reported data in each of the years. Hospitals entering and leaving the study at different times may have introduced bias in that they may have had greatly different characteristics with regards to antibiotic use and resistance. Elimination of hospitals for which we could not identify proportions of resistance in all four years of the study period would have greatly reduced the sample size. Also, the health-systems in this study were from teaching institutions and may not be generalizable to other hospitals or the general population of U.S. health-systems.

Since no diversity index was found to be the “best” in terms of explaining resistance rates in a global fashion, a future study could look at coming up with a new

measure of diversity that may combine some of the characteristics of those in this study. Also, further work could investigate individual hospitals, especially those that have been identified as outliers in this project. There may be certain practices occurring at individual hospitals that better explain some of the measures of resistance seen in this current study. Also, no hospitals were removed from the data set when conducting the analyses of this preliminary study. Identification of hospitals with 'suspect' values and appropriate removal of these observations from the data set could have an important effect on the results of this study. In particular, there may be improvement in the performance of the logistic regression models in terms of fit (lowering the model deviance) and prediction (increasing the percent concordant values, which were relative low for many of the models).

List of drugs used in the study

Classified into fourteen different classes:

1. anti-staphylococcal penicillins (nafcillin, oxacillin, dicloxacillin),
2. 1st & 2nd generation cephalosporins (cefazolin, cefaclor, cefuroxime, cefoxitin, cefotetan, cefixime),
3. 3rd/4th generation cephalosporins (ceftriaxone, cefpodoxime, ceftazidime, ceftidoren, cefdinir, cefprozil, cefotaxime, cefepime, and aztreonam),
4. aminoglycosides (gentamicin, tobramycin, amikacin),
5. beta-lactam/beta-lactamase inhibitors (ampicillin-sulbactam, piperacillin-tazobactam, amoxicillin-clavulanate, ticarcillin-clavulanate),
6. carbapenems (meropenem, imipenem, ertapenem),
7. macrolides (erythromycin, clarithromycin, azitrhomycin),
8. fluoroquinolones (gatifloxacin, moxifloxacin, levofloxacin, ciprofloxacin),
9. sulfamethoxazole/trimethoprim,
10. metronidazole,
11. clindamycin,
12. quinupristin/dalfopristin,
13. vancomycin,
14. linezolid.

Literature Cited

Belsley, D. A., Conditioning Diagnostics, Collinearity and Weak Data in Regression. 1st ed. John Wiley and Sons, Inc., New York, NY, 1991.

Defez C., Fabbro-Peray, P., and Bouziges, N., Risk factors for multidrug-resistant Pseudomonas aeruginosa nosocomial infection. J Hosp Infect. 2004; 57(3):209-16.

Draper, N. R. and Smith, H., Applied Regression Analysis, John Wiley & Sons, Inc., New York, NY, USA, The Correlation between X and Y: 30-35, 1966.

Edwards, A. L., An Introduction to Linear Regression and Correlation, W. H. Freeman, San Francisco, CA, USA, The Correlation Coefficient: 33-46, 1976.

Estridge, B., Basic clinical laboratory techniques. Clifton Park, NY: Thomson Delamar Learning, c2008.

Fortaleza, C. M., Freire, M. P., and Filho, D. C., Risk factors for recovery of imipenem- or ceftazidime-resistant pseudomonas aeruginosa among patients admitted to a teaching hospital in Brazil. Infect Control Hosp Epidemiol. 2006; 27(9):901-06.

Frank, E. and Todeschini, R., The data analysis handbook, Elsevier Science B.V New York, NY:, pp.52-53, 1994.

Greenwood, D., Medical microbiology: a guide to microbial infections: pathogenesis, immunity, laboratory diagnosis and control. Churchill Livingstone, New York, NY, 2007.

Hsu, D. I., Okamoto, M.P., and Murthy, R., Fluoroquinolone-resistant Pseudomonas aeruginosa: risk factors for acquisition and impact on outcomes. J Antimicrob Chemother. 2005;55(4):535-41.

Krebs, C.J., Species Diversity Measures. Ecological Methodology. 1989; 357-360.

Lee, S. O., Kim, N. J., and Choi, S. H., Risk factors for acquisition of imipenem-resistant Acinetobacter baumannii: a case-control study. Antimicrob Agents Chemother. 2004; 48(1):224-8.

MacDougall, C., Harpe, S. E., and Powell, J. P., Pseudomonas aeruginosa, Staphylococcus aureus, and fluoroquinolone use. Emerg Infect Dis. 2005;11(8):1197-204.

Martínez, J., Comparison of antimicrobial cycling and mixing strategies in two medical intensive care units. Crit Care Med 2006; 34: 329-36.

Miles, J. and Shevlin, M., Applying Regression &Correlation, Thousand Oaks, California: SAGE Publications Inc, Building models with regression and correlation: 20-23, 2001.

Myers, R. H., Classical and Modern Regression with Application, Second Edition DUXBURY, Pacific Grove, CA 93950 USA, 1990.

Palmer, K. M. and Young, J. P. W., Higher Diversity of Rhizobium leguminosarum Biovar viciae Populations in Arable Soils than in Grass Soils. App and Envi Micro 2000; 66: 2445-2450.

Powell, J., Antibiotic Diversity and Bacterial Resistance, Virginia Commonwealth University, Richmond, Virginia. 2007.

SAS Institute, Inc. SAS software for windows, Release 9.1.3 Service Pack 3. Cary, N.C.: SAS Institute, Inc., 2003.

Sandiumenge, A., Impact of diversity of antibiotic use on the development of antimicrobial resistance. J. Antimicrob. Chemother. 2006: 57, 1197–1204.

Schwaber, M. J., De-Medina, T., and Carmeli, Y., Epidemiological interpretation of antibiotic resistance studies – what are we missing? Nature Reviews-Microbiology. 2004; 2: 979-983.

Simpson, E. H., Measurement of diversity. Nature 1949; 163:688. (reviewed in http://www.offwell.free-online.co.uk/simpsons.htm).

StataCorp. Stata software for windows, College Station, TX, 2005.

Wax, G., Bacterial resistance to antimicrobials. CRC Press, Boca Raton, FL. 2008.

Zillich, A. J., Antimicrobial use control measures to prevent and control antimicrobial resistance in US hospitals. Infection Control Hosp. Epidemiol. 2006; 27:1088-1095.

VITA

Christine Awuor Ouma was born on December 25, 1981 in Nairobi, Kenya to the parents of Jonathan and Anna Ouma. Christine graduated from Kutztown University of Pennsylvania in may 2005 with a Bachelor of Science degree in Mathematics. From 2006 to 2008, Christine Ouma worked as a graduate teaching assistant at Virginia Commonwealth University, while working on her Master's degree in statistics. Christine is in the process of looking for a job, and plans to work for a year then go back to school in the Fall of 2009 to complete a PhD.